A Lightweight Deep Learning Approach for Spatiotemporal Heat Forecasting

Anonymous Authors[1]

Abstract

Across the world, climate change is getting increasingly worse. This has significant global impacts for living conditions and wellbeing. In particular, extreme heat poses a growing risk to mortality. In 2003, in response to a devastating heatwave, the Spanish Ministry for Health began work on dedicated frameworks to predict and mitigate the public health impacts of extreme heat. Ultimately, these frameworks rely fundamentally on forecasts of daily maximum temperature. Current operational forecast products perform well, particularly those produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). However, their flagship numerical model, IFS, is very computationally expensive and their flagship AI model, AIFS Ensemble, requires large amounts of data. This study evaluates whether a lightweight deep learning approach can provide useful short-range temperature forecasts using atmospheric reanalysis data.

We train a convolutional neural network based on a U-Net architecture on ERA5 reanalysis data to predict daily maximum 2m air temperature over the Iberian Peninsula across multiple forecast lead times. Model performance is examined with respect to three aspects: sensitivity to the length of the input temporal context, the contribution of different atmospheric variables, and comparison with current operational forecast products produced by the European Centre for Medium-Range Weather Forecasts (ECMWF).

Results show that forecast accuracy is strongly dependent on lead time and only weakly affected by extending the input context beyond seven days. Feature analysis indicates that surface pressure creates the most consistent improvements at short lead times. When benchmarked against archived ECMWF forecasts, the proposed model exhibits useful skill at very short horizons but loses accuracy as lead time increases. Despite this, the model trains on substantially less data and a smaller set of environmental variables. It therefore provides a more efficient alternative suitable for deployment on edge devices with minimal computational resources and less than 0.5GB of GPU memory.

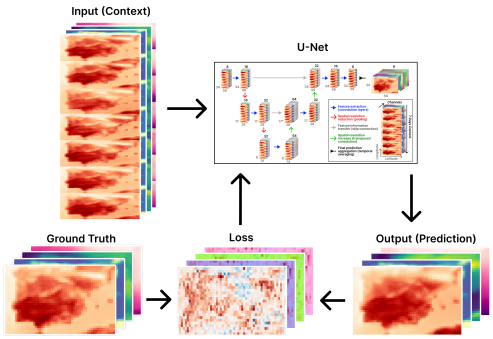

Figure 1. Visualisation of the model training process with the U-Net architecture. The example shown is a model with a 7 day context window and 4 features.

1. Introduction

Extreme heat is widely recognised as one of the most severe and rapidly intensifying consequences of climate change. Recent observational analysis has shown that global warming has not only continued but accelerated since the mid-2010s, with the most recent decade exhibiting the fastest rate of warming observed in the instrumental record (Foster & Rahmstorf). This acceleration is translating into more frequent, longer-lasting, and more intense heat extremes.

Regional climate projections consistently indicate that Southern Europe is especially vulnerable to these changes. Using high-resolution simulations from the EURO-CORDEX project, Lorenzo et al. (2021) projected a *mean increase of 104% in heatwave intensity* across the Iberian Peninsula for the near-future period 2021–2050. This

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

presents a problem due to the associated societal impacts, most notably increased mortality risk. Physiological studies demonstrate that the capacity of the human body to safely dissipate heat diminishes non-linearly with increasing temperature and humidity, particularly for older adults (Vanos et al.). Recent global analyses further show that even under present-day climate conditions, large parts of the world already experience periods where safe physical activity becomes severely constrained due to heat exposure, with these limitations expanding rapidly as temperatures rise (Parsons et al.).

The public health consequences of intensified heat are already evident. Empirical studies across Europe indicate that all-cause and age-specific risks will grow substantially in coming decades. For example, machine-learning-based mortality modelling for Greece projects a sharp rise in heatwave-attributable deaths among elderly populations by the mid-21st century, even under moderate emission scenarios (Petrou et al.). Taken together, these findings highlight the urgent need not only to understand long-term climate trends, but also to anticipate short-term and regional extremes that directly drive health outcomes.

Historically, weather and climate prediction have relied on physically based numerical weather prediction (NWP) and climate models. These models are grounded in first-principles representations of atmospheric physics and have led to major advances in forecasting skill. However, their high computational cost, sensitivity to parameterisations, and limitations in representing local extremes present challenges for applications requiring rapid, high-resolution, or impact-focused forecasts. In parallel with these constraints, the availability of long, high-quality reanalysis datasets has expanded dramatically.

A key contributor to recent climate-data-driven research is the ECMWF Reanalysis v5 (ERA5), which provides globally complete, physically consistent hourly estimates of atmospheric, land, and surface variables from 1940 to the present (Setchell). By assimilating vast quantities of observational data into a single dynamical framework, ERA5 offers a foundation for both physical analysis and data-driven modelling. Such datasets have enabled the increasing adoption of machine learning approaches in climate science.

Machine learning methods are well suited to extreme heat forecasting for two reasons. First, heat extremes arise from non-linear interactions between large-scale circulation, land-atmosphere coupling, and local surface conditions. Second, impacts such as human liveability and mortality respond non-linearly to temperature and humidity, as demonstrated by physiology-based heat stress models (Vanos et al.; Parsons et al.).

Beyond their scientific relevance, improvements in short-term temperature forecasts have direct and practical implications for heat-risk preparedness and response. In Spain, the *National Plan of Preventative Actions for the Effects of Excess Heat on Health*, coordinated by the Spanish Ministry of Health and State Meteorological Agency (AEMET), constitutes the primary framework for mitigating heat-related morbidity and mortality. The Plan explicitly relies on advance forecasts of daily maximum and minimum temperatures over lead times of several days to activate graded health-risk alerts and to trigger preventive measures at national, regional, and local levels. As a result, forecast accuracy and reliability are central to the effectiveness of these interventions, particularly during prolonged or anomalous heat events (Sanidad, 2025).

As extreme heat becomes both more frequent and more intense, the operational demands placed on existing forecast systems are increasing. While current numerical weather forecasting models produced by ECMWF and distributed via AEMET provide the backbone of these early-warning systems, there is growing interest in progressive data-driven approaches that may offer similar results with substantially reduced computational cost.

In this study, we develop a predictive model based on a U-Net CNN architecture to forecast 2m air temperature across the Iberian Peninsula using multi-variable inputs derived from ERA5 reanalysis data. The model is trained to produce spatio-temporal temperature forecasts across Iberian Peninsula. To evaluate performance across different horizons, we train models across a range of forecast lead times. Within this framework, the study addresses the following research questions:

1. How does the length of the input context window influence model performance at different lead times?
2. Which input variables contribute most strongly to model performance?
3. How does the proposed U-Net-based approach compare to current operational forecast products?

Overall, the results show that forecast skill is far more sensitive to lead time than to the length of the input temporal context, with little benefit gained from extending the context window beyond seven days. Feature analysis indicates that *surface pressure* and *10m wind* provide modest model improvements at short lead times, whereas *specific humidity at 850hPa* contributes little in this configuration. When compared against current operational forecast products produced by ECMWF, the proposed U-Net model exhibits useful short-range (D+1, D+2) skill. The model does not match the accuracy or robustness of established operational systems as lead time increases, but generates forecasts at substantially less computational and memory cost.

2. U-Net Convolutional Neural Network

The proposed model adapts a 3D U-Net architecture to the task of short-range temperature forecasting from spatio-temporal reanalysis data. The network processes a sequence of atmospheric states as a unified input then uses convolutional feature extraction to identify spatial patterns within each feature. This enables the model to capture how evolving atmospheric conditions influence future temperature fields, using only a compact set of input variables and limited temporal context.

The model operates on three-dimensional inputs, where each sample is structured as channels $\times$ time $\times$ latitude $\times$ longitude. Each channel corresponds to a distinct atmospheric variable. As illustrated in Figure 1, the network is trained end-to-end by comparing the aggregated prediction to the target field, with optimisation performed using a loss computed equally across all predicted variables. A detailed visualisation of the internal encoder–decoder structure and skip connections is provided in Figure 7.

A key design choice in this work is how temporal information is aggregated. Instead of explicitly predicting a sequence of future states, the network retains the temporal dimension throughout the forward pass and produces multiple internal estimates of the same target field from different parts of the input sequence. These estimates are then combined via simple averaging to produce a single prediction. This effectively allows the model to form a consensus forecast from multiple temporal perspectives, acting as a lightweight form of temporal ensembling.

3. Results

3.1. RQ1 - Context and Lead Time

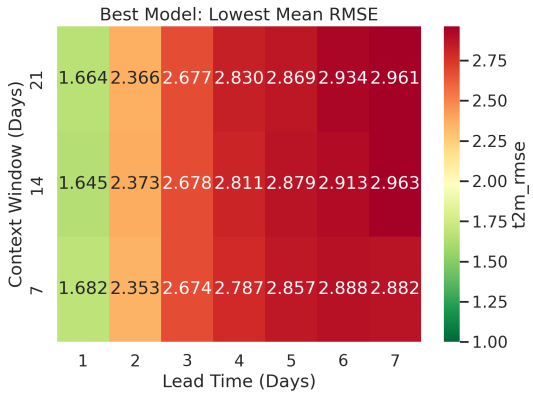

Figure 3. Heatmap showing the best (lowest) mean RMSE scores (see calculation details in Appendix G.1) of each set of models (grid search through hyperparameters) for varying combinations of lead time (D+1 to D+7) and context (D-7 to D-21).

Two patterns emerge from Figures 2 and 3. First, increasing the length of the context window from 7 to 14 or 21 days does not lead to a substantial or systematic improvement in forecast accuracy for lead times up to seven days. In Figure 3, the best mean RMSE values for each lead time are similar across all three context lengths, and no consistent advantage is observed for longer context windows. This indicates that, for the proposed U-Net architecture and variable set, a seven-day context is sufficient to capture the relevant short-term temporal information, and extending the context further yields diminishing returns.

Second, forecast error increases rapidly with lead time beyond approximately three days. Figure 2 shows an upward shift in the distribution of mean RMSE values between D+1 and D+3, followed by a continued increase for longer lead times. While the rate of degradation slows at longer horizons, there is no clear evidence of a performance plateau within the tested range up to D+42. Instead, the spread of errors narrows slightly at longer lead times, suggesting that models converge towards a broadly similar level of skill.

Taken together, these results suggest that the proposed model retains useful predictive skill primarily at short lead times, with D+1 and D+2 forecasts performing substantially better than longer-range forecasts. Beyond D+3, the absolute error becomes large enough that extending the forecast horizon further offers limited practical benefit in terms of accuracy. On the basis of these findings, subsequent experiments in this study focus on short lead times (up to three days) and adopt a seven-day context window as a parsimonious and empirically justified configuration.

3.2. RQ2 - Feature Analysis

Feature relevance was evaluated independently for lead times of one, two, and three days. For each lead time, features are ranked according to their mean relative improvement compared to the *2m temperature*-only baseline. Details of feature rankings are provided in Table 1 for 1, 2 and 3 days lead time. Additionally, the performances for each feature combination for each lead time are provided in Appendix Figure 1.

Across all lead times, mean relative improvement deteriorates as forecast horizon increases, indicating a general loss of exploitable predictive signal. Regarding which auxiliary variables perform best, several consistent patterns emerge. *Surface pressure* is the most robust, providing improvements at all lead times. *10 m wind* and *temperature at 850 hPa* are valuable at shorter horizons but lose effectiveness beyond two days. *Specific humidity at 850 hPa* consistently degrades performance and ranks lowest at every lead time.

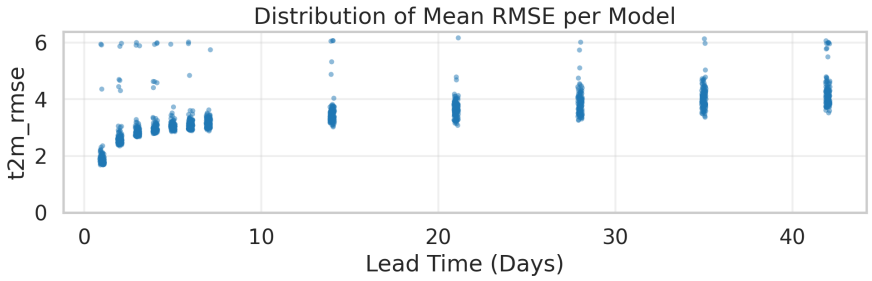

Figure 2. Scatter plot showing the distribution of mean RMSE scores (see calculation details in Appendix G.1) of each set of models (grid search through hyperparameters) for varying lead times: D+1 through to D+42, with fixed context (7 days). A small jitter is added to make individual runs easier to see in dense areas.

Lead Time	Rank	Feature	Mean Relative Improvement
1 Day	1	10m Wind	2.8%
	2	Surface Pressure	2.3%
	3	Temperature at 850 hPa	1.7%
	4	2m Dewpoint Temperature	0.9%
	5	Specific Humidity at 850 hPa	-0.3%
2 Days	1	Surface Pressure	1.7%
	2	10m Wind	0.8%
	3	Temperature at 850 hPa	0.8%
	4	2m Dewpoint Temperature	0.5%
	5	Specific Humidity at 850 hPa	-0.4%
3 Days	1	Surface Pressure	0.3%
	2	10m Wind	-0.2%
	3	2m Dewpoint Temperature	-0.4%
	4	Temperature at 850 hPa	-0.4%
	5	Specific Humidity at 850 hPa	-0.9%

Table 1. Feature rankings and performance impact across one, two, and three day lead times, based on Mean Relative Improvement to the *2m temperature* baseline. Positive values indicate improved performance and negative values indicate degradation.

Lead Time	Model	Evaluation Period	Mean RMSE
1 Day	Proposed U-Net Model	2024–2025	1.931
	ECMWF IFS 0.25°	2024–2025	1.415
	Proposed U-Net Model	2025 only	1.872
	ECMWF AIFS 0.25°	2025 only	1.189
2 Days	Proposed U-Net Model	2024–2025	2.788
	ECMWF IFS 0.25°	2024–2025	1.469
	Proposed U-Net Model	2025 only	2.779
	ECMWF AIFS 0.25°	2025 only	1.251
3 Days	Proposed U-Net Model	2024–2025	2.918
	ECMWF IFS 0.25°	2024–2025	1.562

Table 2. Comparison of the proposed U-Net model (setup details are in Appendix Table 1) against ECMWF forecast baselines across forecast lead times and available evaluation periods. Mean RMSE is reported as the primary performance metric (calculation details are in Appendix G.1).

3.3. RQ3 - Performance vs Operational Baselines

Table 2 shows that, across all lead times and all directly comparable evaluation windows, the archived ECMWF baselines outperform the proposed U-Net model in terms of Mean RMSE. For the one-day forecast, the proposed U-Net model attains a Mean RMSE of 1.931, compared with 1.415 for ECMWF IFS, and 1.872, compared with 1.189 for ECMWF AIFS. The same pattern holds at two days, where the proposed model records 2.788 against 1.469 for IFS, and 2.779 against 1.251 for AIFS. At three days, the proposed model reaches 2.918, compared with 1.562 for IFS. No three-day AIFS archive was openly available for comparison.

There is a clear pattern is that forecast error increases with lead time for all models, but the increase is substantially steeper for the proposed U-Net model than for the ECMWF baselines. Over 2024–2025, the proposed model degrades from 1.931 at one day to 2.788 at two days and 2.918 at three days, whereas the IFS baseline increases only from 1.415 to 1.469 and then 1.562 across the same lead times. Similarly, over the 2025-only comparison, the AIFS baseline remains consistently lower than the proposed model at one and two days. This suggests that while the proposed model retains some predictive skill at short lead times, it does not match the stability or overall accuracy of the operational ECMWF systems as lead time increases.

The appendix tables provide additional context for these comparisons. Appendix Table 2 shows that the same ordering is preserved when Median RMSE, Mean MAE, Median MAE, and the corresponding standard deviations are considered, indicating that the difference is not an artefact of the mean alone. Appendix Table 3 also reports the proposed model over its full 2013–2025 evaluation period, where the same lead-time degradation is visible. Taken together, these results suggest that the proposed U-Net model remains substantially less accurate than ECMWF baselines but achieves useful short-term forecasts with orders-of-magnitude lower computational cost.

Impact Statement

This work aims to develop efficient deep learning methods for short-range temperature forecasting using minimal data and computational resources. By enabling low-cost and scalable forecasting, it can support heat-risk preparedness and decision-making in under-resourced or computationally constrained environments.

References

AIFS: a new ECMWF forecasting system. URL https://www.ecmwf.int/en/newsletter/178/news/aifs-new-ecmwf-forecasting-system.

Who we are. URL https://www.ecmwf.int/en/about/who-we-are.

Temperaturas umbrales de disparo de la mortalidad atribuible al calor en España en el periodo 2000-2009. Instituto de Salud Carlos III. doi: 10.4321/repisalud.5425. URL http://hdl.handle.net/20.500.12105/5425.

AIFS ENS becomes operational. URL https://www.ecmwf.int/en/newsletter/185/earth-system-science/aifs-ens-becomes-operational.

PLAN NACIONAL DE ACTUACIONES PREVENTIVAS DE LOS EFECTOS DEL EXCESO DE TEMPERATURAS SOBRE LA SALUD. 2005.

Implementation of AIFS single v2 - forecast user - ECMWF confluence wiki, 2026. URL https://confluence.ecmwf.int/display/FCST/Implementation+of+AIFS+Single+v2.

Kairós index interactive map and data, 2026. URL https://momo.isciii.es/kairos/.

What is a heatwave?, 2026. URL https://weather.metoffice.gov.uk/learn-about/weather/types-of-weather/temperature/heatwave.

MoMo Interactive Panel, 2026. URL https://momo.isciii.es/panel_momo/.

Previous runs API | open-meteo.com, 2026. URL https://open-meteo.com/en/docs/previous-runs-api.

Asadollah, S. B. H. S., Khan, N., Sharafati, A., Shahid, S., Chung, E.-S., and Wang, X.-J. Prediction of heat waves using meteorological variables in diverse regions of iran with advanced machine learning models. 36 (7):1959–1974. ISSN 1436-3259. doi: 10.1007/s00477-021-02103-z. URL https://doi.org/10.1007/s00477-021-02103-z.

Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J. P., Adeli, E., and Merhof, D. Medical image segmentation review: The success of u-net. 46(12): 10076–10095. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3435571. URL https://ieeexplore.ieee.org/abstract/document/10643318.

Bhoopathi, S., Kumar, N., Somesh, and Pal, M. Evaluating the performances of SVR and XGBoost for short-range forecasting of heatwaves across different temperature zones of india. 24:100204. ISSN 2590-1974. doi: 10.1016/j.acags.2024.100204. URL https://www.sciencedirect.com/science/article/pii/S259019742400051X.

Board, E. Climate change and health: report by the secretariat, 2008. URL https://iris.who.int/items/c5e2d65c-5c81-4ebb-bd82-3ad000ff6c79.

Calvo, J. Modelos numéricos usados en AEMET. modelo del centro europeo y HARMONIE-AROME. URL https://etrp.wmo.int/pluginfile.php/66002/mod_resource/content/1/ModelosNumericosAEMET_JavierCalvo.pdf.

Calvo Sánchez, F. J. Modelos atmosféricos. In Santos Burguete, C. (ed.), *Física del caos en la predicción meteorológica*, pp. 129–143. Agencia Estatal de Meteorología. doi: 10.31978/014-18-009-X.10. URL http://www.aemet.es/documentos/es/conocermas/recursos_en_linea/publicaciones_y_estudios/publicaciones/Fisica_del_caos_en_la_predicc_meteo/10_Modelos_atmosfericos.pdf.

ECMWF. IFS documentation CY49r1 - part v: Ensemble prediction system. URL https://www.ecmwf.int/en/elibrary/81627-ifs-documentation-cy49r1-part-v-ensemble-pr

Foster, G. and Rahmstorf, S. Global warming has accelerated significantly. 53(5):e2025GL118804. ISSN 1944-8007. doi: 10.1029/2025GL118804. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2025GL118804. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2025GL118804.

Gong, B., Langguth, M., Ji, Y., Mozaffari, A., Stadtler, S., Mache, K., and Schultz, M. G. Temperature forecasting by deep learning methods. 15 (23):8931–8956. ISSN 1991-959X. doi: 10.5194/gmd-15-8931-2022. URL https://gmd.copernicus.org/articles/15/8931/2022/.

Islam Sanjid, T., Somadder, A., Ahmed, T., Islam Sanjid, T., Somadder, A., and Ahmed, T. Feasibility assessment of machine learning for predicting heatwaves in bangladesh. 39. ISSN 0187-6236. doi: 10.20937/atm.53438. URL http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S0187-62362025000100023&lng=es&nrm=iso&tlng=en. Publisher: Instituto de Ciencias de la Atmósfera y Cambio Climático, UNAM.

Khan, N., Shahid, S., Ismail, T. B., and Behlil, F. Prediction of heat waves over pakistan using support vector machine algorithm in the context of climate change. 35(7):1335–1353. ISSN 1436-3259. doi: 10.1007/s00477-020-01963-1. URL https://doi.org/10.1007/s00477-020-01963-1.

Kiefer, S. M., Lerch, S., Ludwig, P., and Pinto, J. G. Can machine learning models be a suitable tool for predicting central european cold winter weather on subseasonal to seasonal time scales? 2(4). ISSN 2769-7525. doi: 10.1175/AIES-D-23-0020.1. URL https://journals.ametsoc.org/view/journals/aies/2/4/AIES-D-23-0020.1.xml.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. URL http://arxiv.org/abs/1412.6980.

Lentze, G. ECMWF's ensemble AI forecasts become operational. URL https://www.ecmwf.int/en/about/media-centre/news/2025/ecmwfs-ensemble-ai-forecasts-become-operational.

Lorenzo, N., Díaz-Poso, A., and Royé, D. Heatwave intensity on the iberian peninsula: Future climate projections. 258:105655, 2021. ISSN 01698095. doi: 10.1016/j.atmosres.2021.105655. URL https://linkinghub.elsevier.com/retrieve/pii/S0169809521002076.

Ma, J., Song, J., Guo, J., Fu, Y., Cai, Y., Zhang, Y., Tian, J., and Li, M. Spatial-interaction-context and temporal-attention u-net model for spatio-temporal prediction of sea surface temperature field. 90:104446. ISSN 2352-4855. doi: 10.1016/j.rsma.2025.104446. URL https://www.sciencedirect.com/science/article/pii/S2352485525004372.

Meteorología, A. E. d. Predicción probabilística y postproceso de modelos - agencia estatal de meteorología - AEMET. gobierno de españa. URL https://www.aemet.es/es/idi/prediccion/prediccion_probabilistica.

Modigliani, D. U. The history of ECMWF forecasts and UEFs. 2025. URL https://events.ecmwf.int/event/464/contributions/5296/attachments/3244/5410/UEF2025_Modigliani.pdf.

Parsons, L. A., Baldwin, J. W., Guzman-Echavarria, G., Jay, O., Kalmus, P., Staudmyer, H., Vanos, J. K., and Wolff, N. H. Intensifying global heat threatens livability for younger and older adults. 4(1):015013. ISSN 2752-5309. doi: 10.1088/2752-5309/ae3c3a. URL https://doi.org/10.1088/2752-5309/ae3c3a.

Petrou, I., Kassomenos, P., and Kyriazis, N. Assessing future heatwave-related mortality in greece using advanced machine learning and climate projections. 16(9):1093. ISSN 2073-4433. doi: 10.3390/atmos16091093. URL https://www.mdpi.com/2073-4433/16/9/1093. Publisher: Multidisciplinary Digital Publishing Institute.

Ren, J., Wang, C., Sun, L., Huang, B., Zhang, D., Mu, J., and Wu, J. Prediction of sea surface temperature using u-net based model. 16(7):1205. ISSN 2072-4292. doi: 10.3390/rs16071205. URL https://www.mdpi.com/2072-4292/16/7/1205.

Rodríguez Ballesteros, C. Criterios objetivos para localizar y caracterizar las olas de calor. URL https://repositorio.aemet.es/handle/20.500.11765/6036. Accepted: 2017-02-02T07:09:47Z Company: Agencia Estatal de Meteorología Distributor: Agencia Estatal de Meteorología Institution: Agencia Estatal de Meteorología Label: Agencia Estatal de Meteorología Publisher: Asociación Meteorológica Española.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. URL http://arxiv.org/abs/1505.04597.

Sanidad, M. d. Plan nacional de actuaciones preventivas de los efectos del exceso de temperatura sobre la salud 2025, 2025. URL https://www.sanidad.gob.es/areas/sanidadAmbiental/riesgosAmbientales/calorExtremo/publicaciones/docs/planNacionalExcesoTemperaturas_2025.pdf.

Setchell, H. ECMWF reanalysis v5. URL https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5.

Siddiqui, M. H. S., Abbasi, M. D., Mir, S. Z., Nadeem, G., Majeed, M. K., Khawer, S. K., Younus, I., and Kashif, M. PREDICTION OF MAXIMUM AIR TEMPERATURE FOR DEFINING HEAT WAVE IN VARIOUS STATES OF PAKISTAN USING MACHINE

LEARNING ALGORITHM. pp. 942–960. ISSN 3007-3138. URL https://thesesjournal.com/index.php/1/article/view/321.

Straaten, C. v., Whan, K., Coumou, D., Hurk, B. v. d., and Schmeits, M. Using explainable machine learning forecasts to discover subseasonal drivers of high summer temperatures in western and central europe. 150(5):1115–1134. ISSN 1520-0493, 0027-0644. doi: 10.1175/MWR-D-21-0201.1. URL https://journals.ametsoc.org/view/journals/mwre/150/5/MWR-D-21-0201.1.xml.

Tobias, A., Armstrong, B., Zuza, I., Gasparrini, A., Linares, C., and Diaz, J. Mortality on extreme heat days using official thresholds in spain: a multi-city time series analysis. 12(1):133. ISSN 1471-2458. doi: 10.1186/1471-2458-12-133. URL https://doi.org/10.1186/1471-2458-12-133.

V, R. C., Johnvictor, A. C., and N, P. S. Comparative analysis of machine learning approaches for heatwave event prediction in india. 15(1):22431. ISSN 2045-2322. doi: 10.1038/s41598-025-04634-9. URL https://www.nature.com/articles/s41598-025-04634-9. Publisher: Nature Publishing Group.

Vanos, J., Guzman-Echavarria, G., Baldwin, J. W., Bongers, C., Ebi, K. L., and Jay, O. A physiological approach for assessing human survivability and liveability to heat in a changing climate. 14(1):7653. ISSN 2041-1723. doi: 10.1038/s41467-023-43121-5. URL https://www.nature.com/articles/s41467-023-43121-5.

Wang, G. and Ren, T. Design of sports achievement prediction system based on u-net convolutional neural network in the context of machine learning. 10(10). ISSN 2405-8440. doi: 10.1016/j.heliyon.2024.e30055. URL https://www.cell.com/heliyon/abstract/S2405-8440(24)06086-9.

Weirich-Benet, E., Pyrina, M., Jiménez-Esteve, B., Fraenkel, E., Cohen, J., and Domeisen, D. I. V. Subseasonal prediction of central european summer heatwaves with linear and random forest machine learning models. 2(2). ISSN 2769-7525. doi: 10.1175/AIES-D-22-0038.1. URL https://journals.ametsoc.org/view/journals/aies/2/2/AIES-D-22-0038.1.xml.

Wu, X., Hong, D., and Chanussot, J. UIU-net: U-net in u-net for infrared small object detection. 32: 364–376. ISSN 1941-0042. doi: 10.1109/TIP.2022.3228497. URL https://ieeexplore.ieee.org/abstract/document/9989433.

Zhu, Y., Zhi, X., Lyu, Y., Zhu, S., Tong, H., Mamtimin, A., Zhang, H., and Huo, W. Forecast calibrations of surface air temperature over xinjiang based on u-net neural network. 10. ISSN 2296-665X. doi: 10.3389/fenvs.2022.1011321. URL https://www.frontiersin.org/journals/environmental-science/articles/10.3389/fenvs.2022.1011321/full.

Çiçek, , Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In Ourselin, S., Joskowicz, L., Sabuncu, M. R., Unal, G., and Wells, W. (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pp. 424–432. Springer International Publishing. ISBN 978-3-319-46723-8. doi: 10.1007/978-3-319-46723-8_49.

A. Background

A.1. Heatwaves vs. Extreme Heat

A.1.1. Meteorology vs. Epidemiology

The Spanish State Meteorological Agency (AEMET) *"consider a heatwave an episode of at least three consecutive days, in which 10% [of the 137] weather stations register maximum temperature above the 95th percentile of the series of daily maximum temperatures for the months of July and August during the period 1971-2000"* (Rodríguez Ballesteros) (Author's translation).

This is similar to the Met Office's definition (met, 2026), which consider a heatwave a period of at least three consecutive days with daily maximum temperatures exceeding the temperature threshold, which is calculated using the 1991-2020 climatology of daily maximum temperature on the 15 July.

Both agencies calculate the threshold for each geographic region, with the AEMET calculated for each of its 137 weather stations, and the Met Office calculating for each of the 92 UK counties.

However, for practical purposes, the Spanish Government do not use the AEMET definition.

In 2003, an unprecedented European heatwave resulted in approximately 44,000 deaths (Board, 2008), highlighting the urgent need for heat-health prevention strategies. In response, the Spanish Ministry for Health launched the *National Plan for Preventive Actions Against the Effects of Excessive Temperatures on Health* in 2004 (noa, 2005). Its core mechanism relies on defining temperature thresholds and assigning risk levels based on those thresholds to guide public health interventions.

In the Plan, the authors explain that the meteorological definition doesn't account for public health impacts. They emphasise a need for an epidemiological definition for heatwaves that factors in health impacts, particularly mortality. As part of this, they argue that the need to have consecutive days is too strict, as short episodes of extreme heat can still have notable health impacts, and temperature thresholds should be calculated based on the point in which extreme heat creates an anomaly in mortality (Sanidad, 2025).

A.1.2. Calculating the Temperature Threshold

In the first National Plan (2005), regions were defined as provincial capitals, each linked to a single weather station. Thresholds were calculated using the 95th percentile of historical maximum and minimum temperatures, with minor exceptions in some regions. Between 2005 and 2014, this percentile-based approach remained standard. However, in 2012, Tobias et al. (Tobias et al.) highlighted inconsistencies in these thresholds, showing they did not uniformly reflect mortality risk, relative to the thresholds, across regions.

Responding to this, in 2015, the first thresholds to take mortality into account were introduced. *"Trigger"* temperatures were derived from mortality and temperature data spanning 2000–2009 (noa). From 2024 onward, the system became more granular. In place of the previous 52 provincial capitals, 182 *"meteosalud"* zones were defined. In terms of epidemiological threshold calculation, thresholds now incorporate more complicated methods developed in an extended study, split between two papers, by López-Bueno, Gil, et al. The study uses more sophisticated statistical methods to analyse the relationship between mortality and temperature, and uses this as its baseline in setting maximum temperature thresholds.

A.1.3. Assigning the Associated Risk Level

Risk levels were initially determined by counting consecutive days where predicted maximum and minimum temperatures exceeded their thresholds. From 2004 to 2022, the scale ranged from Level 0 (no days) to Level 3 (five consecutive days above thresholds).

In 2023, a new algorithm introduced a new calculation based on temperature deviations over three days, weighted by daily risk factors:

$$RiskScore = \sum_{d=1}^{3} (Tmax_d - Tthreshold) * RiskFactor_d$$

Risk Levels: 0 (=0), 1 (> 0, ≤ 3.5), 2 (> 3.5, ≤ 7), 3 (> 7), as seen in Table 3.

Risk Score	Risk Level	Denomination
0	0	Absence of risk
$> 0, \leq 3.5$	1	Low risk
$> 3.5, \leq 7$	2	Medium risk
> 7	3	High risk

Table 3. Definition of heatwave risk levels for health (Sanidad, 2025).

Note: currently the risk factor for every region is 1, which the National Plan says will change with time. However, it has stated this since the introduction of the algorithm in 2023.

A.1.4. Public Health Response to Heat-Related Risk Levels

Within the National Plan, the above risk factors are used to trigger progressively more intensive public health responses (Sanidad, 2025). Lower risk levels correspond to baseline preventive action, centred on public information, continuous surveillance, and coordination between health, social, and meteorological services. As risk levels increase, responses shift towards active risk management, including heightened communication with healthcare providers and social services, targeted protection of vulnerable populations, and strengthened institutional coordination at regional and national scales. At the highest risk level, health authorities evaluate the need for additional emergency-oriented actions.

A.1.5. This Study

This study adopts the epidemiological perspective outlined above by treating heat as a continuous variable rather than a categorical event. Accordingly, the forecasting task is framed as a regression problem, in which daily maximum 2m air temperature is predicted directly. This approach allows temperature thresholds and health-related risk metrics to be applied downstream if required, while avoiding hard-coded assumptions about heatwave duration or severity. In the present work, no thresholds or heatwave classification are applied; instead, the focus is on evaluating the accuracy and robustness of daily maximum temperature forecasts themselves.

A.2. Excess Heat to Excess Deaths

A.2.1. MoMo: Monitoring Heat-Related Mortality in Spain

Spain's Daily Mortality Monitoring System (MoMo), first developed alongside the *National Plan for Preventive Actions Against the Effects of Excessive Temperatures on Health* in 2004, detects temperature-related deviations in observed daily deaths from the expected baseline.

MoMo uses three datasets: mortality records from the National Institute of Statistics (INE), temperature data from the State Meteorological Agency (AEMET), and population statistics by age, sex, and province from INE.

Using these inputs, MoMo applies a Generalized Additive Model (Poisson GAM) to estimate excess deaths. The simplified explanation is that the model first predicts mortality based on time and temperature variables, then recalculates using just the time variable. The difference between these two estimates represents deaths attributable to temperature.

The key variables calculated by the system are: base estimated deaths (expected deaths based on historical time-series), notified deaths (raw counts actually received from civil registries), observed deaths (an adjusted series that accounts for reporting delay to better approximate actual deaths), deaths attributed to deficient temperatures (excess cold), deaths attributed to excess temperatures (excess heat).

The system operates at multiple scales: national, regional (autonomous communities), and provincial. Data can be split into sex and age group, which could provide utility for investigations into how heat affects specific groups. Results are updated daily and publicly accessible through an interactive dashboard, which visualises observed deaths, baseline estimates, and temperature-attributable mortality. Figure 4 shows this. Note the seasonal mortality patterns, and the spikes in deaths attributable to excess heat, like in August 2025.

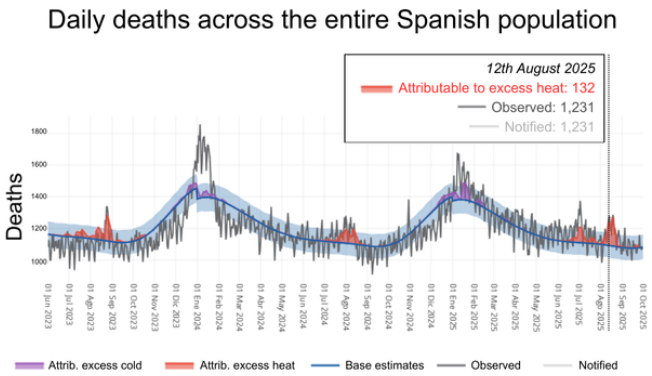

Figure 4. Time series graph from the MoMo interactive panel (mom, 2026) showing daily death statistics within the Spanish population. The graph highlights the 12th August 2025, where there were 1231 observed deaths, with 132 attributable to excess heat - more than 10%.

Probability	Kairós Index	Mortality Risk
<0.4	1	Small or null
$\geq 0.4, \leq 0.6$	2	Moderate
>0.6	3	High

Table 4. Definition of Kairós Indices based on the probability of a $\geq$10% increase in mortality attributable to excess heat.

(Sanidad, 2025)

A.2.2. Kairós: Anticipating Mortality Risk from Extreme Heat

Alongside the Daily Mortality Monitoring System is the Kairós alert system CITE Kairos, designed to predict mortality risk associated with excess heat.

The system is based around the Kairós Index, which estimates the probability of a $\geq$10% increase in mortality attributable to excess heat over the current day and the next four days. This probability is derived from the same Poisson GAM framework used by MoMo, used with real-time meteorological forecasts. Based on this probability, Kairós assigns one of three risk indices, shown in Table 4.

These alerts are geographically broken down to the provincial level and can also be split by age and sex. Figure 5 shows an example of an interactive map from the MoMo website.

A.2.3. This Study

The evidence outlined above demonstrates a clear and well-established pathway linking excess ambient heat to increased mortality risk and the mobilisation of public health responses. Central to this pathway is the reliance on short-term forecasts of daily maximum temperature, which act as the primary upstream driver of epidemiological assessment and intervention planning. This aim of this study is to evaluate the robustness of a novel deep learning based temperature prediction method, providing a foundational input that supports downstream heat-health surveillance and decision-making systems.

A.3. European Centre for Medium-Range Weather Forecasts

The European Centre for Medium-Range Weather Forecasts (ECMWF) is an intergovernmental organisation supported by European Member and Co-operating States, tasked with producing global numerical weather predictions and Earth-system data products for its members (ecm). The Spanish State Meteorological Agency (AEMET) uses ECMWF output as a fundamental component of its operational forecasting system, integrating ECMWF global forecasts into national prediction, post-processing, and decision-support workflows (Calvo; Meteorología).

Weather prediction has developed through a series of methodological and technological shifts. Early forecasting relied heavily on human interpretation of observations and empirical rules, before the growth of electronic computing made

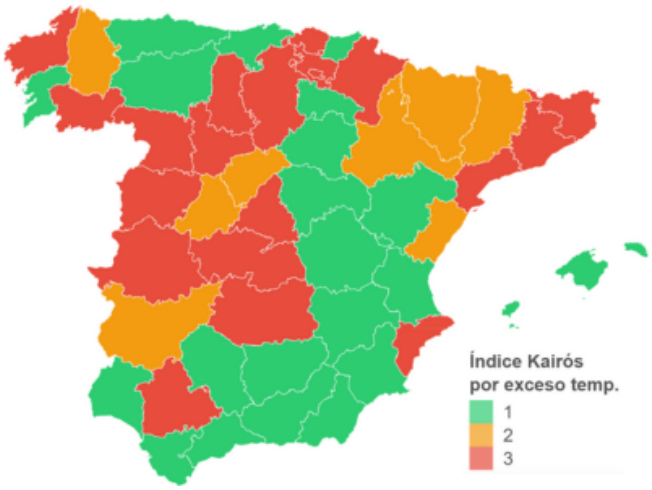

Figure 5. Screenshot of the interactive map of the Spanish Peninsula and Balearic Islands, showing the Kairós Index for each province. Green indicates Kairós Index 1, orange 2, red 3. The map shows the 12th August 2025, where large parts of Spain registered Kairós 3 alerts, indicating high mortality risk during a severe heatwave. (kai, 2026)

numerical weather prediction operationally feasible. In the ECMWF context, experimental operational forecasting began in 1979, with fully operational daily forecasting following in 1980 (Modigliani, 2025).

In 1987, the ECMWF began development on the Integrated Forecasting System (IFS), which became operational in 1994 (Modigliani, 2025). The IFS consolidated many of the components needed in an operational forecasting centre into a single forecasting framework, and over time it was extended with increasingly advanced data assimilation, higher spatial resolution, coupling to ocean and wave models, and operational ensemble forecasting (Modigliani, 2025; ECMWF). ECMWF's first operational ensemble forecasts appeared in the early 1990s, and ensemble methods have since become central to modern forecasting because they explicitly represent uncertainty in the initial state and model formulation rather than producing only a single deterministic scenario (Modigliani, 2025; ECMWF).

These developments greatly improved forecast quality, but they also made operational forecasting increasingly dependent on large-scale computing infrastructure. Numerical weather prediction remains computationally expensive because it requires repeated solution of high-dimensional, non-linear equations, together with data assimilation, post-processing, and verification (Calvo Sánchez; Calvo). As model resolution increases, the computational cost grows rapidly, while uncertainty does not disappear and often needs to be handled through ensemble systems and statistical post-processing (Calvo; Calvo Sánchez). In operational practice, this means that numerical forecasts are not simply read off directly from the model. Instead, they are interpreted by trained forecasters and frequently complemented by probabilistic products and statistical correction methods (Calvo Sánchez; Meteorología).

More recently, the field has entered a new phase in which machine learning and artificial intelligence are being explored. In January 2024, the ECMWF introduced the Artificial Intelligence Forecasting System (AIFS), which is based data-driven methods. By 2025, both AIFS Single and AIFS Ensemble had entered operations (aif; Modigliani, 2025); the latter was said to *"generate forecasts over 10 times faster than the physics-based forecasting system, while reducing energy consumption by approximately 1,000 times."* (Lentze). The AIFS uses a graph-based encoder-decoder together with a transformer-style processor, reflecting the wider move towards neural architectures designed for high-dimensional spatio-temporal prediction. ECMWF reports that AIFS produces highly skilled forecasts for upper-air and surface variables, while IFS remains the established operational forecasting backbone. The current situation is therefore best understood as the coexistence of physics-based and data-driven forecasting systems, rather than a complete replacement of one by the other (aif; 2026; ECMWF).

This study is situated within this transition. Rather than attempting to reproduce the full complexity of a modern operational system such as IFS or AIFS, it will investigate whether a single deep learning architecture, a U-Net convolutional neural network, can learn useful predictive structure from historical ERA5 data for the Iberian Peninsula. The goal is to assess whether such a model can provide a competitive and computationally efficient approach to predicting daily maximum temperature, and then to compare its performance against archived ECMWF forecast products used as operational baselines.

A.4. Machine Learning for Temperature Prediction

A growing body of literature has explored the use of machine learning methods for predicting extreme heat and high-temperature conditions, motivated by the limitations of traditional statistical and dynamical approaches in capturing non-linear atmospheric processes. Early work in this area focused predominantly on the classification of heatwave events, often defined using fixed temperature thresholds and duration criteria. Examples include studies applying support vector machines, random forests, and gradient boosting to identify heatwave days several days in advance across South Asia and the Middle East (Khan et al.; Asadollah et al.; Bhoopathi et al.). While these approaches demonstrated encouraging skill, their dependence on discrete heatwave definitions can limit flexibility when transferring results across regions or impact frameworks.

More recent studies have shifted attention towards predicting temperature directly, framing the problem as a regression task. Siddiqui et al. (Siddiqui et al.) applied multiple machine learning regression models to predict daily maximum air temperature across regions of Pakistan using reanalysis and satellite-derived variables, explicitly motivated by heatwave preparedness. Similarly, Gong et al. (Gong et al.) provided a comprehensive assessment of deep learning approaches for temperature forecasting, demonstrating that neural networks can outperform conventional baselines across a range of temporal scales while capturing complex spatial and temporal dependencies.

In parallel, research has increasingly targeted Europe and longer forecasting horizons. Van Straaten et al. (Straaten et al.) employed explainable machine learning techniques to identify key subseasonal drivers of high summer temperatures in Central and Western Europe, highlighting the role of large-scale circulation patterns. Related work by Weirich-Benet et al. (Weirich-Benet et al.) and Kiefer et al. (Kiefer et al.) demonstrated that machine learning models can provide skilful predictions of temperature extremes and anomalous conditions on sub-seasonal timescales, although forecast skill decreases rapidly at shorter leads if local processes are not well represented.

Despite these advances, most machine learning studies surveyed in Europe have focused either on sub-seasonal variability or on the prediction of categorical extreme events, rather than short-term daily maximum temperature. By contrast, several recent studies in South Asia have demonstrated the feasibility of short-range heat prediction using reanalysis-driven machine learning models. V et al. (V et al.) and Islam Sanjid et al. (Islam Sanjid et al.) evaluated a wide range of classifiers for next-day and three-day-ahead heatwave prediction, identifying surface temperature, humidity, pressure, and wind fields as consistently important predictors.

Building on these insights, the present study adopts a direct temperature forecasting approach, focusing on short lead times relevant for public health preparedness. By predicting daily maximum temperature as a continuous variable using ERA5 data, this work aligns with recent methodological trends while addressing a gap in the European literature for short-term, high-resolution heat prediction using modern machine learning architectures.

A.5. U-Net Convolutional Neural Networks

The U-Net convolutional neural network architecture was originally introduced by Ronneberger et al. (Ronneberger et al.) for biomedical image segmentation. Its defining encoder–decoder structure, combined with skip connections between corresponding resolution levels, enabled precise localisation while maintaining contextual awareness, even when trained on relatively small datasets. This design subsequently became highly influential in medical imaging, with extensions such as the 3D U-Net adapting the framework to volumetric data through three-dimensional convolutions (Çiçek et al.). The effectiveness and versatility of U-Net have since been extensively documented within the medical imaging literature (Azad et al.).

Beyond biomedical applications, U-Net architectures have been adapted to a wide range of tasks and disciplinary contexts. For example, variants incorporating attention mechanisms, dense connections, or nested structures have been proposed for object detection and prediction problems in domains as varied as sports performance analysis (Wang & Ren) and infrared small object detection (Wu et al.). These adaptations reflect a broader shift in which U-Net is treated not solely as a segmentation model, but as a general-purpose architecture for learning structured spatial representations.

More recently, U-Net-based models have been extended to spatio-temporal prediction tasks in environmental and geoscientific applications. Studies have demonstrated their effectiveness for forecasting gridded temperature fields, either as stand-alone prediction models or as post-processing tools for numerical forecasts, by integrating temporal dynamics through recurrent layers or attention mechanisms (Zhu et al.; Ren et al.). Advanced architectures further incorporate explicit spatial interaction contexts and temporal attention to model complex dependencies in sea surface temperature fields (Ma et al.).

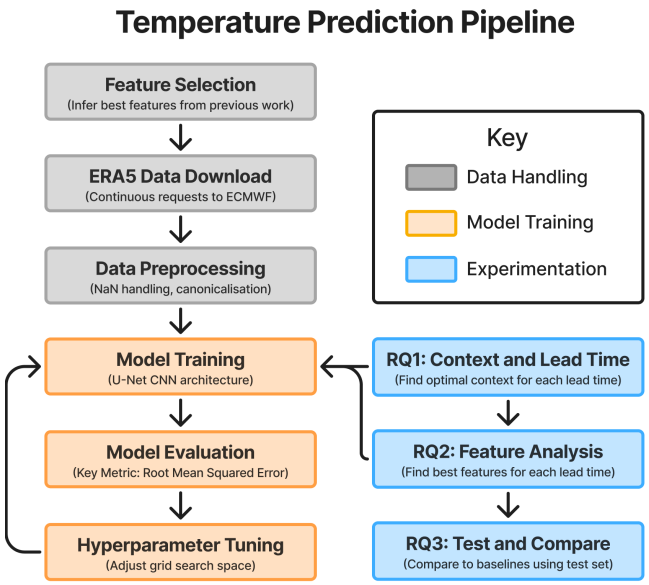

Figure 6. Workflow for building and experimenting on the prediction model. The pipeline is split into three phases: data handling, model training and experimentation. Experiments investigating the impact of the context window length and feature set refer back to the model training phase as many models had to be trained to investigate their effects. The data handling phase only happened once.

Atmospheric Variable	Symbol	Temporal Resolution	Level
Temperature	t2m	Daily Max	2m
Dewpoint Temperature	d2m	Daily Mean	2m
Surface Pressure	sp	Daily Mean	Surface
U-Component of Wind	u10	Daily Mean	10m
V-Component of Wind	v10	Daily Mean	10m
Specific Humidity	sh850	Daily Mean	850hPa
Temperature	t850	Daily Mean	850hPa

Table 5. Details of the atmospheric variables used in this study. All were downloaded for May, June, July and August from 1940-2025, within the spatial region of 35.75°S–44°N and -10°W–3.75°E at a spatial resolution of 0.25° × 0.25°.

In this study, U-Net is adopted within this latter context, leveraging its proven capacity to learn structured spatial representations while extending its application to short-term temperature prediction relevant to heat-related risk assessment.

B. Temperature Prediction System

B.1. Pipeline

The system pipeline, shown in Figure 6, demonstrates the iterative workflow used for data handling, model training and experimentation. First, an initial set of input features is inferred based on prior machine-learning heatwave and temperature prediction studies, subject to data availability and practical constraints. Atmospheric variables are then sequentially downloaded from the ECMWF Climate Data Store using ERA5 reanalysis products.

All downloaded NetCDF files are combined into a single, unified dataset. During this stage, missing values are handled, variables are canonicalised, and units are standardised. Once preprocessing is complete, the dataset is split into training, validation and testing subsets.

The model is trained on the training subset and evaluated against ground-truth data from the validation subset, using root

mean squared error (RMSE) as the primary performance metric. Model training, evaluation, and hyperparameter tuning form an iterative loop: the hyperparameter grid search space is adjusted to identify configurations that minimise validation loss, after which the model is retrained and re-evaluated.

Following convergence of the training pipeline, the workflow branches into the chronological series of experiments. In the first experiment, temporal context length is varied to examine its relationship with prediction lead time. Then, feature subsets are evaluated empirically to assess their impact on predictive performance. Finally, selected model configurations are tested and compared against baseline approaches using held-out test data.

B.2. Dataset

This study uses atmospheric reanalysis data from ERA5, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 provides a spatially and temporally complete representation of the atmospheric state, making it well suited for data-driven forecasting applications that require long, homogeneous input records. ERA5 has been widely adopted in recent machine-learning-based heatwave and temperature prediction studies due to its resolution, consistency, and availability across multiple decades (Weirich-Benet et al.; Khan et al.; Asadollah et al.; Kiefer et al.; Straaten et al.).

The atmospheric variables used in this work are listed in Table 5. Daily maximum 2m temperature (t2m) is the primary prediction target and is always included as an input feature. All remaining variables are treated as auxiliary predictors. Due to download size and speed limits of ERA5 data, the number of atmospheric variables we could work with was severely limited. As such, we prioritised a compact collection of commonly used predictors. Predictors chosen were based on their repeated empirical use in prior machine-learning heatwave studies and their availability as stable, gridded reanalysis fields (Khan et al.; Asadollah et al.; Weirich-Benet et al.; Siddiqui et al.; Islam Sanjid et al.; Bhoopathi et al.; V et al.; Kiefer et al.; Gong et al.; Straaten et al.).

All variables were extracted at the native ERA5 spatial resolution of $0.25° \times 0.25°$. The spatial domain spans 35.75°S to 44°N and -10°W to 3.75°E, covering all of the Iberian Peninsula, the region of interest for this study. Data were downloaded for the period 1940–2025. Primarily due to the ERA5 download limits, we chose to focus on boreal summer (June, July and August), as it contains the highest concentration of extreme heat events, and we include May in our dataset to provide sufficient temporal context for models using longer context windows. For example, to predict June 1st with a 14 day context window would require the data from May 18th to May 31st. This seasonal restriction is consistent with prior heatwave prediction studies using machine-learning approaches (Khan et al.; Weirich-Benet et al.).

B.3. U-Net Convolutional Neural Network

The U-Net is a convolutional neural network architecture designed to learn patterns in spatial data by compressing an input to capture high-level structure and then expanding it back to its original resolution to recover fine-scale detail. The compression stage (encoder) allows the model to understand broad spatial relationships, while the expansion stage (decoder) reconstructs detailed outputs using information preserved from earlier steps.

The U-Net in this study operates on sequences of gridded atmospheric fields defined on a regular latitude–longitude grid. Each input sample consists of an n-day context window (n = 7, 14, or 21), where each day contains multiple features (in this study, we use various subsets of the environmental variables defined in Table 5). These variables are represented as input channels, such that each channel corresponds to one physical field sampled over the spatial domain. As data progress through the encoder, spatial resolution is reduced from the native 34×56 grid to 17×28 and then 8×14 via spatial pooling, while the number of channels increases. This is done by transitioning from raw environmental variables to learned feature channels, where each channel encodes a distinct spatial pattern or physical relationship inferred from combinations of the original features.

Convolutional layers transform input channels into progressively richer feature representations by combining information from neighbouring grid points across latitude, longitude, and time. Early layers learn local spatial structures tied closely to individual environmental variables, while deeper layers encode broader, multi-variable patterns representative of macro-scale atmospheric behaviour. Spatial downsampling increases the effective spatial footprint of these features, while the temporal dimension is preserved throughout the network, allowing features to integrate information across the full context window. The decoder restores spatial detail by fusing large-scale contextual features with fine-resolution information preserved from the encoder (skip connections).

Finally, the network produces a temporally distributed prediction, retaining the input time dimension throughout the forward

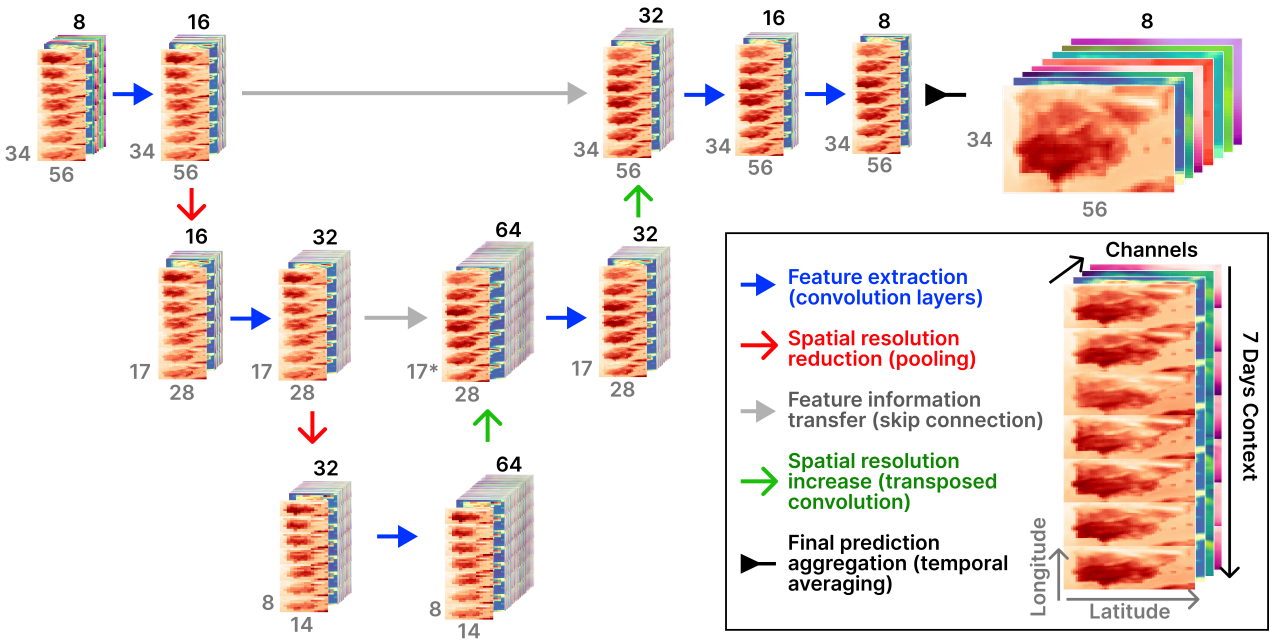

Figure 7. U-net architecture used in this study. The example shown is a model with a 7 day context window and 8 initial features. The number of channels is denoted on top of the box. The number of grid points in each spatial direction are denoted at the bottom of the box. *After the first upscaling step, the longitude is 16 points, so a padding process is required to restore it back to 17 points. This process is not included in the diagram.

pass. Each temporal slice represents a separate estimate of the same target field (e.g. a 1-day lead time), based on different parts of the input sequence. These estimates are then averaged across the time dimension to produce a single final prediction.

An example of how we use the U-Net CNN architecture within this study is visualised in Figure 7.

To train a prediction model using this architecture, the network receives an n-day (7, 14, or 21) input sequence and produces a corresponding set of spatial predictions. These predictions are temporally aggregated into a single forecast field, which is compared against the ground-truth observation for the target day. As shown in the training diagram, the resulting loss is propagated back through the network to update model parameters. The loss is computed as the mean squared error across all predicted variables simultaneously, such that each contributes equally to the optimisation objective.

This process is illustrated in Figure 1. Not shown is early stopping terminating training when validation loss across all predicted variables stops improving. The model training process repeated for different lead times (1, 2, and 3-day leads are used in this study) as each lead time requires training different models, as, during the training process, the lead time is handled by shifting the target day relative to the input window.

B.4. Implementation and Experimental Details

All experiments were implemented in Python using the PyTorch framework, with NumPy, pandas, and xarray used for numerical operations and NetCDF-based data handling. Training, evaluation, and analysis were fully scripted in Jupyter notebooks, with identical code paths used across all experimental settings.

A single 3D U-Net architecture was used throughout. Architectural configuration, loss function, and optimisation strategy were fixed. The input feature set, context length, lead time, and selected hyperparameters varied between runs. Models were optimised using the Adam optimiser (Kingma & Ba) with mean squared error loss and early stopping based on validation loss.

The downloaded ERA5 NetCDF files were combined into a single NetCDF file with consistent physical units. Training and validation splits were applied strictly by year; 1940-2000, and 2001-2012 respectively. 2013-2025 was left exclusively for the testing stage.

Hyperparameter	Lead Time vs Context	Feature Analysis
Learning Rate	6e-4, 3e-4, 1e-4, 6e-3, 3e-3, 1e-3	3e-3, 6e-3
Weight Decay	0.0, 5e-6, 1e-5, 6e-5	0.0, 5e-6, 1e-5
Batch Size	4, 8, 16, 32, 64	4, 8
Epoch Limit	15	50
Patience	3	3

Table 6. Hyperparameters and search ranges used in grid-search optimisation for the Lead Time vs Context and Feature Analysis experiments.

Normalisation was applied to all input and target variables to stabilise optimisation and ensure comparability across variables with different physical units and dynamic ranges. For each variable, the mean and standard deviation were computed once using only the training subset and then fixed for all experiments. Inputs were normalised by subtracting the training mean and dividing by the training standard deviation, with the inverse transformation applied during evaluation.

To reduce wall-time without affecting numerical results, several GPU-oriented optimisations were employed. Automatic mixed-precision training was enabled, data loading was optimised using pinned memory, prefetching, and persistent workers, and cuDNN autotuning was enabled. All optimisations were applied uniformly across experiments and do not affect reproducibility or reported metrics. More detail on how these optimisations work is included in Appendix D.

All experiments were executed within a Docker container on fixed hardware consisting of 4 AMD Ryzen 5955WX CPUs, an NVIDIA RTX 5000 Ada GPU, and 32 GB RAM. This setup exceeds the computational requirements of the model but ensured that hardware variability did not confound experimental comparisons.

B.5. Hyperparameter Tuning

The hyperparameters used to train the model vary depending on the experimental contexts, detailed in Section 3.

In all cases, a grid search was conducted. The specific hyperparameters used are detailed in Table 6. In the case of the Lead Time vs Context experiment, the purpose of the search was both to find the combination of hyperparameters that resulted in the lowest loss (illustrated in Figure 3), and to provide a rich dataset to chart patterns with more certainty (illustrated in Figure 2). In the case of feature analysis, and subsequently picking the best overall model, the purpose of the search was solely to find the combination of hyperparameters that resulted in the lowest loss (illustrated in Appendix Figure **??**).

The set of hyperparameters used in the feature analysis was derived from analysing the data from the Lead Time vs Context experiment and removing the most ineffective hyperparameters.

C. Experimental Setups

C.1. RQ1 - Context and Lead Time

RQ1 investigates how forecast performance varies as a function of (i) the length of the historical context window provided to the model and (ii) lead time. All experiments in this section use the same U-Net convolutional architecture described in Section B.3, with a fixed set of input variables (*2m temperature, 2m dewpoint temperature, surface pressure, 10m wind*) and spatial domain. Models were trained using a grid search over learning rate, weight decay, and batch size (see Table 6) for each combination of lead time and context length.

Lead time experiments span forecast horizons from D+1 to D+42 days, with context fixed to seven days for the long-range analysis. For the joint context–lead analysis, lead times from D+1 to D+7 were combined with context windows of 7, 14, and 21 days. Each trained model produces daily maximum 2m temperature forecasts, which are evaluated against ERA5 ground truth using the daily spatial RMSE metric defined in Appendix G.1. For each configuration, performance is summarised using the mean RMSE across all evaluation days, and distributions across hyperparameter runs are retained to

assess robustness.

Figure 2 shows the distribution of mean RMSE values across all runs for each lead time with fixed context, while Figure 3 shows the best-performing model for each combination of lead time and context window. A corresponding heatmap using the median run for each grid search is provided in Appendix 1.

C.2. RQ2 - Feature Analysis

The objective of this analysis is to assess the empirical contribution of atmospheric variables to model performance. All feature analyses were conducted using a seven-day input context and evaluated independently for lead times of one, two, and three days. As it is the target variable, *2m temperature* is included in every model configuration. All other variables are treated as optional auxiliary features.

The auxiliary variables considered are *2m dewpoint temperature*, *surface pressure*, *temperature at 850 hPa*, *specific humidity at 850 hPa*, and *10m wind*. *10m wind* is treated as a single feature, represented jointly by its zonal (u) and meridional (v) components.

For each lead time, an exhaustive set of feature combinations was evaluated. A dedicated hyperparameter grid search (see Table 6 for details) was performed separately for every feature combination, and the best-performing configuration was retained based on mean RMSE (see Appendix G.1 for how this is calculated). To assess feature utility, all model configurations were compared against a baseline model trained using only *maximum 2m temperature*.

Performance is evaluated using a relative improvement metric defined with respect to this baseline. This formulation, detailed in Appendix G.2, provides a normalised measure of how much each feature configuration reduces error relative to the temperature-only model. Positive values of relative improvement indicate a reduction in RMSE and therefore improved predictive skill, while negative values indicate that the inclusion of additional variables degrades performance compared to the baseline.

Lead times were analysed separately to avoid conflating feature relevance with forecast horizon effects. Per-lead-time feature rankings and tables showing the top-performing feature combinations for each lead time are reported in Appendix F.

C.3. RQ3 - Performance vs Operational Baselines

The comparison in RQ3 evaluates the best-performing proposed U-Net model at each lead time against archived ECMWF forecast products. While the development of the physics-based IFS and data-driven AIFS was outlined in Section A.3, here they serve as operational baselines. Within AEMET's workflow, ECMWF output is central to national forecasting, alongside numerical model HARMONIE-AROME, which provides higher-resolution short-range guidance, whereas ECMWF products cover longer forecast ranges (Calvo; Meteorología). However, HARMONIE-AROME was not included in the retrospective comparison due to the lack of an openly accessible, reproducible archive; AEMET primarily provides interactive visualisations rather than downloadable historical data. Consequently, ECMWF IFS and AIFS archives were selected as the most practical open-source operational baselines.

For each lead time, the proposed U-Net model made predictions from the held-out test data. The setup of each model is listed in Appendix Table 1. These prediction files were produced as NetCDF outputs and evaluated over one full evaluation window for the proposed model, and two period-matched windows for direct comparison against the ECMWF baselines. This structure allows the proposed model to be assessed both over its full test set and over the maximum periods directly comparable to each ECMWF baseline.

The ECMWF baseline archives were downloaded using the Open-Meteo Previous Runs API (ope, 2026). In this study, archived forecasts were retrieved for ECMWF IFS 0.25° and ECMWF AIFS Single 0.25° at one, two, and three-day lead times. Because a single request did not cover the full Iberian study region, the domain was downloaded in two spatial tiles and then concatenated into a single gridded dataset for each model and lead time. The downloaded products were available at hourly resolution and were converted to daily maximum 2m temperature before being saved as NetCDF files.

To ensure a consistent comparison, both the proposed-model forecasts and the ECMWF archives were aligned to the ground-truth NetCDF using the shared coordinates `valid_time`, latitude, and longitude.

Variable names were standardised across files and all forecast temperatures were converted to a common unit (Celsius) prior to evaluation. Where forecast and truth grids did not coincide exactly, the truth field was interpolated to the forecast grid

using nearest-neighbour interpolation. Evaluation was restricted to summer (June–August), consistent with the study's focus on summertime extreme heat.

After alignment, daily spatial errors were computed over the full study domain and summarised over time according to the metric definitions given in Appendix G.1. Mean RMSE is reported as the primary headline metric in this paper, while Median RMSE, Mean MAE, Median MAE, and the corresponding standard deviations are reported in the appendix. Full-period metrics for the proposed U-Net model over 2013–2025 are reported separately in Appendix Table 3, while the period-matched baseline comparison is given in Appendix Table 2.

Finally, the available baseline archives differed by model family: usable IFS archive data were available for 2024–2025, while usable AIFS archive data were available only for 2025, with no valid three-day AIFS archive available from Open-Meteo. Consequently, Table 2 reports results over the maximum available comparison period for each model and lead time.

D. GPU Optimisation Methods

Automatic mixed-precision (AMP) training reduces computational cost by performing selected operations in reduced numerical precision while maintaining higher precision where required for numerical stability. In this work, AMP was implemented using `torch.amp`, with BF16 used when supported by the hardware and FP16 otherwise. This approach accelerates training and reduces memory usage without changing model architecture, optimisation targets, or evaluation metrics.

Data loading performance was improved through several standard techniques. Pinned (page-locked) host memory was enabled to accelerate asynchronous data transfers from CPU to GPU. Prefetching was used to overlap data loading with model computation, reducing GPU idle time. Persistent DataLoader workers were maintained across epochs to avoid repeated worker initialisation overhead, which is particularly relevant for long-running training jobs.

Finally, cuDNN autotuning was enabled to allow the backend to select the most efficient convolution algorithms for the given input shapes and hardware configuration. This optimisation affects only kernel selection and has no impact on numerical correctness or determinism of the training process. All optimisations were applied consistently across experiments to ensure fair comparisons and reproducible results.

E. Context and Lead Time

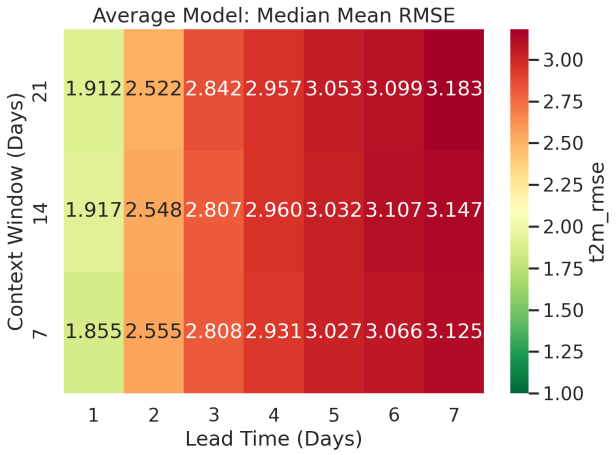

Figure 1. Heatmap showing the average (median) of the mean RMSE scores (the 2m temperature RMSE is calculated for every day of the validation set, then the mean is taken) of each set of runs (grid search through hyperparameters) for varying combinations of lead time (D+1 to D+7) and context (D-7 to D-21).

F. Feature Analysis

Feature combination performance was evaluated independently for each forecast lead time using a fixed input context of seven days. For a given lead time, all possible combinations of the selected atmospheric variables were considered, with *2m temperature* included in every configuration. For each feature combination, a dedicated hyperparameter grid search was performed, and the configuration yielding the lowest validation root mean squared error (RMSE) was retained.

Performance was quantified relative to a baseline model trained using only *2m temperature*. For each feature combination, the difference in mean RMSE between the selected configuration and the baseline model (ΔRMSE) was computed (see Appendix G.3). Negative ΔRMSE values indicate improved predictive performance relative to the baseline, while positive values indicate degraded performance. Feature combinations are ranked according to their ΔRMSE for each lead time, and the full rankings are visualised in Figure **??**.

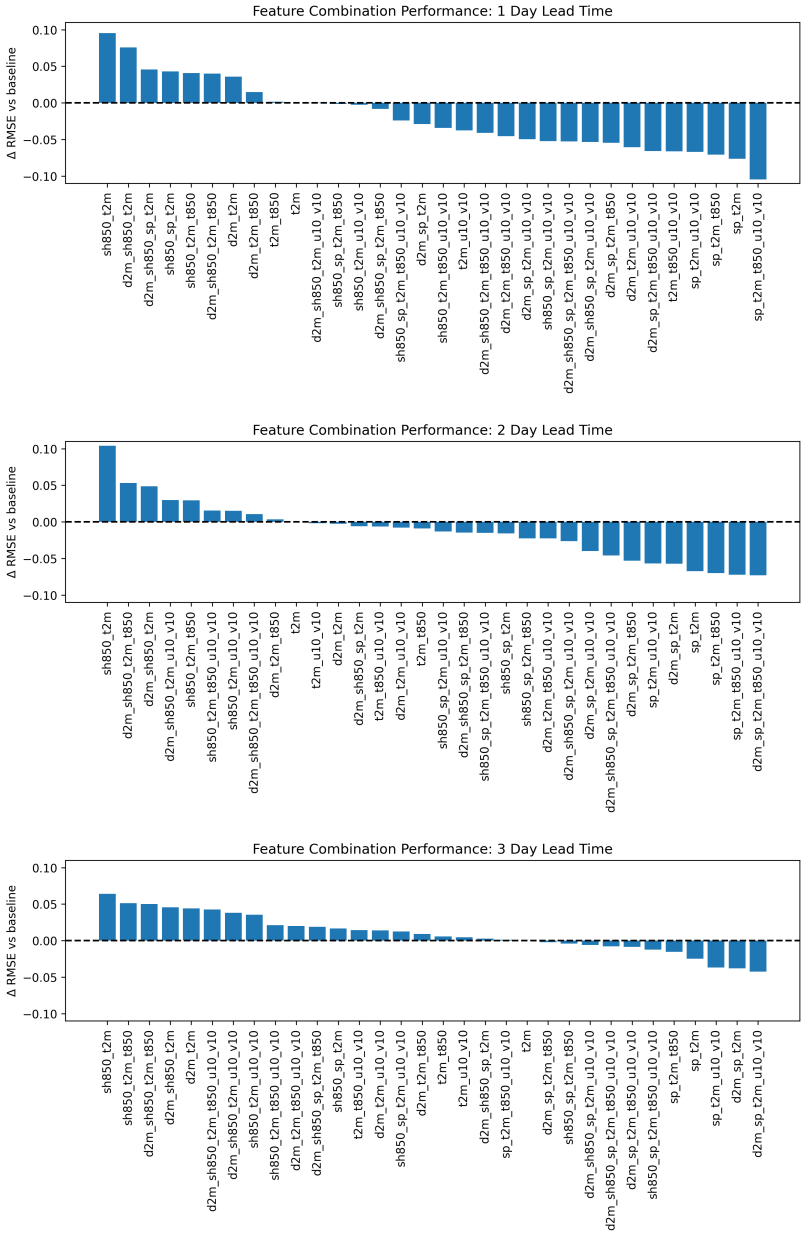

Figure 1. Prediction performance for different input variable combinations at one, two, and three day lead times, expressed as ΔRMSE relative to the *2m temperature* baseline. Negative values indicate improved performance. Axis labels use abbreviated variable symbols for brevity; full variable names are provided in Table 5.

Lead Time	Rank	Feature	Mean ΔRMSE	Median ΔRMSE
1 Day	1	10m Wind	-0.047	-0.051
	2	Surface Pressure	-0.039	-0.050
	3	Temperature at 850 hPa	-0.029	-0.037
	4	2m Dewpoint Temperature	-0.016	-0.035
	5	Specific Humidity at 850 hPa	0.004	-0.002
2 Days	1	Surface Pressure	-0.041	-0.043
	2	10m Wind	-0.019	-0.014
	3	Temperature at 850 hPa	-0.018	-0.015
	4	2m Dewpoint Temperature	-0.013	-0.011
	5	Specific Humidity at 850 hPa	0.009	0.002
3 Days	1	Surface Pressure	-0.009	0.007
	2	10m Wind	0.006	0.008
	3	2m Dewpoint Temperature	0.011	0.011
	4	Temperature at 850 hPa	0.012	0.008
	5	Specific Humidity at 850 hPa	0.023	0.020

Table 1. Feature rankings and performance impact across one, two, and three day lead times, based on mean and median ΔRMSE relative to the *2m temperature* baseline. Negative values indicate improved performance.

G. Performance Metrics

G.1. Mean RMSE

For each day evaluated d, the model produces a spatial prediction for *2m temperature*, which is compared against the corresponding ground-truth field over the spatial region. Let $\hat{y}_{d,i}$ denote the predicted value at grid cell i on day d, and let $y_{d,i}$ denote the corresponding observed value. If the domain contains M grid cells, the daily root mean squared error (RMSE) and daily mean absolute error (MAE) are computed as

$$\mathrm{RMSE}_d = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\hat{y}_{d,i} - y_{d,i}\right)^2}, \tag{1}$$

$$\mathrm{MAE}_d = \frac{1}{M}\sum_{i=1}^{M}\left|\hat{y}_{d,i} - y_{d,i}\right|. \tag{2}$$

These daily values are stored in the `daily_losses.csv` files. Let $\mathcal{D}$ denote the set of evaluation days.

The headline metric reported in the main paper, *Mean RMSE*, is therefore not a single global RMSE computed over all spatial and temporal points simultaneously. Instead, it is the arithmetic mean of the daily spatial RMSE values:

$$\overline{\mathrm{RMSE}} = \frac{1}{N}\sum_{d\in\mathcal{D}} \mathrm{RMSE}_d. \tag{3}$$

Similarly, the *Mean MAE* is defined as the arithmetic mean of the daily spatial MAE values:

$$\overline{\mathrm{MAE}} = \frac{1}{N}\sum_{d\in\mathcal{D}} \mathrm{MAE}_d. \tag{4}$$

To provide robust supplementary summaries in the appendix, the median daily RMSE and median daily MAE are also reported:

$$\widetilde{\mathrm{RMSE}} = \mathrm{median}\left\{\mathrm{RMSE}_d : d \in \mathcal{D}\right\}, \tag{5}$$

$$\widetilde{\mathrm{MAE}} = \mathrm{median}\left\{\mathrm{MAE}_d : d \in \mathcal{D}\right\}. \tag{6}$$

In addition, the spread of daily errors can be quantified using the standard deviation of the retained daily RMSE and MAE values:

$$\sigma_{\mathrm{RMSE}} = \sqrt{\frac{1}{N-1}\sum_{d\in\mathcal{D}}\left(\mathrm{RMSE}_d - \overline{\mathrm{RMSE}}\right)^2}, \tag{7}$$

$$\sigma_{\mathrm{MAE}} = \sqrt{\frac{1}{N-1}\sum_{d\in\mathcal{D}}\left(\mathrm{MAE}_d - \overline{\mathrm{MAE}}\right)^2}. \tag{8}$$

Accordingly, the main paper reports $\overline{\mathrm{RMSE}}$ and $\overline{\mathrm{MAE}}$ as the primary deterministic error metrics for each lead time, while $\widetilde{\mathrm{RMSE}}$, $\widetilde{\mathrm{MAE}}$, and the corresponding standard deviations are provided in the appendix as supplementary measures of robustness and spread.

G.2. Relative Improvement Metric

To complement the absolute error metrics, a relative performance measure is used to quantify the contribution of additional atmospheric variables. Let $\overline{\mathrm{RMSE}}_i$ denote the mean RMSE of a model trained with feature set i, and let $\overline{\mathrm{RMSE}}_{\text{baseline}}$ denote the mean RMSE of the baseline model using only *2m temperature*.

For each feature set, the *relative improvement* is defined as:

$$Relative\ Improvement = \frac{\overline{\mathrm{RMSE}}_{\text{baseline}} - \overline{\mathrm{RMSE}}_i}{\overline{\mathrm{RMSE}}_{\text{baseline}}}. \tag{9}$$

This metric represents the fractional reduction in RMSE relative to the baseline model, such that positive values indicate improved performance and negative values indicate degradation.

To obtain feature-level scores, the relative improvement is averaged over all model configurations in which a given feature appears. For a feature f, let $\mathcal{I}_f$ denote the set of feature sets containing f, with $|\mathcal{I}_f|$ elements. The *Mean Relative Improvement* is therefore defined as:

$$Mean\ Relative\ Improvement = \frac{1}{|\mathcal{I}_f|}\sum_{i\in\mathcal{I}_f}\frac{\overline{\mathrm{RMSE}}_{\text{baseline}} - \overline{\mathrm{RMSE}}_i}{\overline{\mathrm{RMSE}}_{\text{baseline}}}. \tag{10}$$

This value is reported as a percentage.

G.3. ΔRMSE

For completeness, the analysis also makes use of the absolute change in RMSE relative to the baseline model. Let $\overline{\mathrm{RMSE}}_i$ denote the mean RMSE of a model trained with feature set i, and let $\overline{\mathrm{RMSE}}_{\text{baseline}}$ denote the corresponding baseline mean RMSE using only *2m temperature*. The change in error is defined as:

$$\Delta\mathrm{RMSE}_i = \overline{\mathrm{RMSE}}_i - \overline{\mathrm{RMSE}}_{\text{baseline}}. \tag{11}$$

This quantity represents the absolute difference in predictive error introduced by the inclusion of additional variables. Negative values of $\Delta\mathrm{RMSE}_i$ indicate an improvement relative to the baseline, while positive values indicate degraded performance.

In the appendix (Table 1), both the mean and median values, obtained by averaging over all model configurations in which a given feature appears, of $\Delta\mathrm{RMSE}$ are reported for each feature. While relative improvement offers a scale-independent measure of proportional change, $\Delta\mathrm{RMSE}$ preserves the absolute magnitude of the error difference.

Lead Time	Learning Rate	Weight Decay	Batch Size	Environmental Variables
1 Day	0.0003	5e-6	8	Surface pressure, 2m air temperature, temperature at 850hPa, 10m wind.
2 Days	0.0003	5e-6	8	2m dewpoint temperature, surface pressure, 2m air temperature, temperature at 850hPa, 10m wind.
3 Days	0.0003	1e-5	4	2m dewpoint temperature, surface pressure, 2m air temperature, 10m wind.

Table 1. Best-performing model configuration selected for each forecast lead time. All models used a 7-day context window.

Lead Time	Model	Evaluation Period	Mean RMSE	Median RMSE	Mean MAE	Median MAE	RMSE SD	MAE SD
1 Day	Proposed U-Net Model	2024–2025	1.931	1.814	1.495	1.406	0.511	0.401
	ECMWF IFS 0.25°	2024–2025	1.415	1.406	1.014	1.005	0.118	0.090
	Proposed U-Net Model	2025 only	1.872	1.843	1.459	1.416	0.459	0.373
	ECMWF AIFS 0.25°	2025 only	1.189	1.151	0.891	0.865	0.155	0.116
2 Days	Proposed U-Net Model	2024–2025	2.788	2.706	2.160	2.070	0.810	0.658
	ECMWF IFS 0.25°	2024–2025	1.469	1.438	1.056	1.041	0.134	0.100
	Proposed U-Net Model	2025 only	2.779	2.653	2.171	2.060	0.791	0.662
	ECMWF AIFS 0.25°	2025 only	1.251	1.227	0.944	0.914	0.175	0.134
3 Days	Proposed U-Net Model	2024–2025	2.918	2.837	2.271	2.193	0.955	0.773
	ECMWF IFS 0.25°	2024–2025	1.562	1.520	1.130	1.106	0.205	0.152

Table 2. Full comparison of the proposed U-Net model against ECMWF forecast baselines across lead times and available evaluation periods. Metrics are computed using the appendix G.1 definitions based on daily spatial errors summarised over time.

H. Performance vs Baseline

I. Conclusions

I.1. Discussion

This study set out to evaluate whether a U-Net convolutional neural network trained on ERA5 reanalysis data can learn useful predictive structure for short-term daily maximum temperature forecasting over the Iberian Peninsula. The results demonstrate that such a model is capable of producing skilful forecasts at very short lead times, but that its performance is strongly constrained by forecast horizon and does not approach that of contemporary operational forecasting systems as lead time increases.

With respect to temporal context, the findings indicate that forecast accuracy is far more sensitive to lead time than to the amount of historical input provided to the model. Extending the input context window beyond seven days yields little systematic improvement in performance for lead times up to one week. This suggests that, within the proposed architecture and variable set, the dominant short-term predictive signal is contained within recent conditions, and that longer context windows introduce diminishing returns. This result has practical implications for model design, as it supports the use of relatively short input sequences that reduce computational cost without sacrificing predictive skill.

Feature analysis further clarifies how the model exploits physical information. Large-scale dynamical variables, most notably *surface pressure* and *10m wind,* consistently improve performance at short lead times. These variables appear to provide useful information about evolving synoptic-scale conditions that modulate short-term temperature variability. In

Lead Time	Model	Evaluation Period	Mean RMSE	Median RMSE	Mean MAE	Median MAE	RMSE SD	MAE SD
1 Day	Proposed U-Net Model	2013–2025	1.797	1.714	1.391	1.322	0.531	0.414
2 Days	Proposed U-Net Model	2013–2025	2.554	2.418	1.977	1.856	0.823	0.658
3 Days	Proposed U-Net Model	2013–2025	2.783	2.593	2.161	1.987	0.995	0.805

Table 3. Full metrics of the proposed U-Net model across lead times and the full 2013-2025 evaluation period. Metrics are computed using the appendix G.1 definitions based on daily spatial errors summarised over time.

contrast, *specific humidity at 850 hPa* did not provide a clear benefit in this configuration and often degraded performance.

When benchmarked against archived ECMWF forecast products, the proposed U-Net model demonstrates limited competitiveness. Although it retains some predictive skill at very short horizons (D+1 and D+2), it is consistently outperformed by both the physics-based IFS and the data-driven AIFS systems across all comparable lead times. Moreover, forecast error for the proposed model increases more rapidly with lead time than for the ECMWF baselines. These results reinforce the current operational reality: modern numerical and hybrid forecasting systems remain substantially more accurate and robust for deterministic daily maximum temperature prediction.

A key factor underlying this performance gap is the substantial difference in model capacity and input representation. The best-performing models in this study contain approximately 330,805 parameters (five-feature configuration) and 331,254 parameters (six-feature configuration), which is several orders of magnitude smaller than the ECMWF AIFS ensemble, which consists of approximately 229 million parameters (ope). However, in this study, we tested archived data from AIFS Single, not AIFS Ensemble. Data for AIFS Single isn't available, but based on the reported ensemble size of 51 members (Lentze), we estimate ~ 4.5 million parameters for an individual AIFS Single model, which remains substantially larger than the models developed here. In addition to this disparity in parameter count, the models in this study operate on a limited set of at most seven atmospheric variables at daily temporal resolution, whereas AIFS Single incorporates a far richer representation, using well over 100 environmental variables at a finer 6-hourly temporal resolution (aif, 2026).

Computational constraints further distinguish the proposed approach from operational systems. The U-Net models developed in this study require less than 0.5GB of GPU memory during training, meaning that the models will require less than 0.5GB at inference time. In contrast, the ECMWF report that their AIFS-ENS-1.0 model *"requires 38GB of GPU memory by default"* at inference time. No data was available for the AIFS Single model, so a like-for-like comparison cannot be made.

I.2. Limitations and Future Work

Practical constraints associated with ERA5 data access represented one of the most significant limitations of this study. Download limits restricted both the number of atmospheric variables that could be incorporated into the model and the scope of the feature analysis performed. As a result, the feature combinations evaluated were drawn from a relatively small subset of potentially relevant environmental variables, limiting the breadth of conclusions that could be drawn about variable importance. In addition, all experiments were conducted using daily aggregated fields, whereas short-term temperature forecasting may benefit from higher frequency inputs at hourly, three-hourly, or six-hourly resolution. These constraints reflect practical data-access limitations rather than methodological ones and point to clear opportunities for more flexible future analyses.

An additional limitation of the present work is the focus on air temperature alone. A growing body of literature highlights that metrics relevant to heat-related health impacts depend not only on temperature but also on atmospheric moisture, particularly relative humidity, which strongly influences human heat stress and physiological tolerance (Vanos et al.). Consequently, future models aiming to support public-health applications should consider jointly predicting temperature and humidity-related variables, enabling downstream derivation of heat-stress or thermal-comfort indices that more directly reflect physiological risk.

Further work is also required to extend the analysis beyond domain-wide averages. Developing regional or sub-regional evaluation methods would allow performance to be compared more directly with operational baselines in a geographically meaningful way and could reveal spatial heterogeneity in predictive skill across the Iberian Peninsula. Identifying and analysing environmentally distinct sub-regions may help clarify where data-driven approaches are most effective and where their limitations are most pronounced.

Finally, alternative machine learning and deep learning architectures could be explored for the same forecasting task, including transformer-based models, recurrent architectures, or hybrid approaches that explicitly incorporate temporal attention mechanisms. In parallel, ensemble approaches may improve robustness and allow uncertainty to be quantified in ways that are more relevant for decision making contexts.