Journal Name

Crossmark

RECEIVED
dd Month yyyy

REVISED
dd Month yyyy

RESEARCH PAPER

Assessing and Explaining Temporal Deep Learning Models for Wildfire Danger Prediction

Pauline Becker[1,*], Carolina Natel[3], and Peer Nowack[1,2]

[1]Institute of Theoretical Informatics, Chair for AI in Climate and Environmental Sciences, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Institute of Meteorology and Climate Research - Atmospheric Trace Gases and Remote Sensing (IMKASF), Karlsruhe Institute of Technology, Karlsruhe, Germany
[3]Institute of Meteorology and Climate Research - Atmospheric Environmental Research (IMKIFU), Karlsruhe Institute of Technology, Garmisch-Partenkirchen, Germany
*Corresponding author

E-mail: pauline.anne.becker@gmail.com

Keywords: Wildfires, Mediterranean, Explainable AI, SHAP, Transformers, AI in Climate Science

Abstract

Accurate and robust wildfire danger prediction is critical for mitigating the detrimental impacts of fires on ecosystems, public health, and the economy. While Machine Learning (ML) has emerged as a powerful approach to model the complex interactions driving wildfires, their "black-box" nature often limits interpretability, creating a critical trade-off between predictive skill and physical plausibility for trustworthy risk assessment. In this study, we systematically assess the predictive performance and physical consistency of seven temporal deep learning (DL) models against a Random Forest (RF) baseline for next day wildfire danger prediction in the Mediterranean. We apply explainaible AI (xAI) methods to interpret model attributions and assess their alignment with established fire science. Results show that all DL models outperform the RF baseline (F_1-score > 0.8), with Transformer models achieving the highest predictive accuracy by effectively capturing long-range temporal dependencies. However, xAI analyses reveal a key trade-off, in which the higher-performing DL models exhibit lower physically consistency in their averaged driver relationships compared to the simpler RF model. We further investigate how Transformers generated individual wildfire danger predictions through case studies of two similar large fire events in Spain, one correctly predicted (true positive) and one missed (false negative). The analysis demonstrates how differences in driver representation and feature importance can lead to divergent predictions, such as correctly identifying a heatwave-driven event but missing a lightning-induced ignition. Together, these investigations provide a structured evaluation of a wide range of DL models in terms of their predictive accuracy and physical consistency, offering guidance for future wildfire danger forecasting in fire-prone regions, such as the Mediterranean.

1 Introduction

Wildfires were long considered carbon neutral due to vegetation regrowth offsetting emissions [1, 2]. However, this balance is increasingly disrupted by anthropogenic climate change and changes in land use and management practices. In particular, there is a growing global risk of conditions conducive to more frequent, intense, and widespread fires [2, 3, 4, 5]. Beyond their implications for carbon emissions [6, 4, 7, 8], aerosol radiative forcing, and even short-term weather changes [9], wildfires also pose major risks to ecosystems and biodiversity [10], public health [11], water quality and infrastructure [12]. Considering the projected intensification of climatic changes over this century, which is expected to further exacerbate fire weather conditions in many world regions [13], and the inherent complexity of modeling wildfire risk [14, 15, 16], there is a pressing need to improve wildfire predictive capabilities to help develop effective adaptation [17, 18] and mitigation strategies (e.g., through fuel management) [19].

In this context, machine learning (ML) has emerged as a powerful approach, often surpassing traditional process-based fire models, which tend to over-predict high fire danger and produce false

alarms—particularly in fuel-limited biomes [20, 21, 22]. Recent studies show that widely used ML algorithms such as Decision Trees, Random Forests (RFs), XGBoost, and Artificial Neural Networks (ANNs) outperform conventional fire weather indices globally [21] and in regional (e.g., over the Western US) high-resolution wildfire prediction tasks [23]. ML methods can leverage large, heterogeneous datasets to capture complex and non-linear interactions among diverse fire drivers, including climate, vegetation, topography, and human-related factors [1, 24]. To further explore this potential, a few benchmark datasets for wildfire activity have been developed. Such datasets are critical to progress in data-driven wildfire modeling, as they provide standardized frameworks for testing and comparing diverse ML architectures. For instance, the SeasFire data cube [25], with a 0.25° horizontal resolution, has been used to consistently evaluate multiple deep learning (DL) models, including architectures such as U-Net [1] and Vision Transformer [26], for forecasting global burned area patterns across multiple temporal windows at coarse spatial resolutions. Kondylatos et al. [27], in turn, introduced the Mesogeos fire cube [27], covering the entire Mediterranean region at substantially higher spatial resolution (1 km×1 km) and benchmarked first DL models on this dataset, after Kondylatos et al. [28] had previously demonstrated that DL models outperform the Fire Weather Index in predicting next-day wildfire danger in Greece.

Despite the overall rapid progress, many open questions remain before ML and DL methods can be reliably applied in operational forecasting contexts. For example, understanding how model complexity influences model accuracy and interpretability is increasingly recognized as a central challenge in the application of ML DL to forecasting. Di Giuseppe et al. [21], for instance, systematically evaluated multiple ML architectures to assess how model complexity and input data quality affect predictive skill, concluding that higher complexity does not necessarily lead to improved performance. However, their work did not explore whether temporal DL models could provide additional benefits. Complementarily, Li et al. [23] benchmarked ML models against classical approaches and employed explainable AI (xAI) techniques to reveal substantial structural differences among models—even when predictive accuracies were comparable. Increasingly, researchers across environmental modeling disciplines emphasize that purely statistical performance improvements are insufficient. There is a growing demand for models that are not only accurate but also transparent and consistent with scientific intuition in their reasoning.

Finally, we highlight that most data-driven wildfire modeling studies to date have been limited to a small number of specific types of ML models (typically one; in rare cases up to four different approaches) and have lacked systematic, comprehensive intercomparisons of DL methods in terms of both accuracy and explainability [28, 23]. Applications of xAI in wildfire prediction remain scarce, and when present, they have typically only investigated simpler architectures such as Feedforward Neural Networks [29] or RF [30], in combination with one selected xAI method. Furthermore, existing analyses have often focused on specific countries or subregions-such as Turkey [31], France [32], Italy [30], Australia [29], Greece [28], and China [33], which limits the generalizability of their findings.

To address these gaps, we present a comprehensive evaluation of DL methods for wildfire danger prediction across the entire Mediterranean region. We expand upon initial models from Kondylatos et al. [27] by systematically benchmarking seven advanced temporal DL architectures — including Long Short-Term Memory (LSTM) [34] and Temporal Fusion Transformers [35] — against a Random Forest baseline [36]. Our work is further distinguished by its integrated use of xAI techniques to assess not only predictive performance but also model plausibility, so as to more holistically inform future DL wildfire modeling exercises. The three core objectives are: to determine if increased model complexity yields higher performance; to assess its impact on explainability; and to provide insights into the trade-offs between model accuracy and explainability in fire danger predictions. To elucidate the driver attributions behind correct and incorrect predictions, we complement our analysis with two illustrative case studies.

2 Data and Methods

2.1 Data and Preprocessing

In this study, we used the Mesogeos data cube [27] for model training and evaluation. This dataset provides pre-processed daily observational and reanalysis data for key variables characterizing wildfire activity and its drivers at a 1 km×1 km spatial resolution from 2006–2022 across the Mediterranean region. Meteorological variables include surface air temperature, wind speed, wind direction, dewpoint temperature, surface pressure, relative humidity, precipitation, and surface solar radiation, derived from the ERA5-Land dataset [37]. Vegetation status and land surface conditions were represented using daytime and nighttime land surface temperature [38], the Normalized Difference Vegetation Index (NDVI) [39], and the Leaf Area Index (LAI) [40] from MODIS,

alongside soil moisture estimates from the European Drought Observatory (EDO) [41]. Indicators of human presence and activity, such as population density and proximity to roads, were obtained from WorldPop [42]. Terrain characteristics, including elevation and slope, were incorporated using the COP-DEM dataset [43], while land cover classifications were sourced from the Copernicus Climate Change Service [44]. Burned areas were retrieved from EFFIS, while ignition points and ignition dates were estimated using the MODIS Active Fire product [45]. A full list of all Mesogeos predictor variables, their abbreviations, categories, and units is provided in Table S1.

We preprocessed the data following the Mesogeos definitions [27], in which positive samples were defined as fire events exceeding an area of 30 ha around an ignition point. Ignition locations were approximated by computing the spatial centroid of each polygon of burned grid cells, with the nearest grid cell to each centroid designated as the ignition source. For each positive sample, we extracted a 30-day temporal window of predictor variables spanning from day $(t - 30$ to $t - 1)$, excluding the ignition day t. Inputs included all dynamic (e.g., meteorological) and static (e.g., altitude) Mesogeos features, with static layers repeated across the temporal dimension to match dynamic variables. Negative samples were drawn from regions located at least 62 km away from any recorded fire-occurrence radius to minimize the likelihood of selecting unburned locations that nonetheless exhibited high fire danger [27]. To address the pronounced class imbalance, we sampled twice as many negative samples as positives ones, following standard practices [27, 46, 28, 33]. The temporal distribution of negatives was roughly that of the positives. This prevents over-representation of low-fire periods (e.g., winter months), and preserve the seasonal structure of fire occurrence. Missing values, mainly arising from satellite data gaps (e.g., cloud cover), were imputed using the feature-specific temporal mean computed across each sample's entire time series.

2.2 Model Setup

In order to train, evaluate and interpret the models, we implemented a ML pipeline (Figure 1) that included data preprocessing, training and validation (with hyperparameter optimization and early stopping to prevent overfitting), testing, and xAI analysis. Hyperparameters were tuned using GridSearch [47, 48, 23, 49] and Optuna [50], exploring parameters such as learning rate, batch size, dropout rate, and weight decay. The final selected values for each model are reported in Table S3.

We systematically compared seven DL architectures, including Multilayer Perceptron (MLP) [51], Long Short-Term Memory (LSTM) [34], Gated Recurrent Unit (GRU) [52], Convolutional Neural Network (CNN) [53], Transformer [54], Gated Transformer Network (GTN) [27], and Temporal Fusion Transformer (TFT) [35], against a RF [36] baseline model. This comparison was motivated by the need to evaluate whether increasing model complexity translates into superior predictive power for wildfire danger, or whether simpler approaches may already be sufficient. Each architecture entails specific strengths and limitations. RFs capture non-linear relationships through ensembles of decision trees but lack temporal awareness [55]. MLPs, grounded in classical feed-forward neural networks, are compact and computationally efficient in capturing non-linear interactions in flattened inputs but do not incorporate temporal or spatial structure [56]. CNNs, originally devised for image classification [53], can be adapted to multivariate time-series forecasting by reshaping sequences into pseudo-images (time as "height," features as "width"), allowing shared-weight convolutions to extract local temporal and spatio-temporal patterns, but they remain limited in capturing long-range dependencies [53]. LSTMs introduce cell states and gating mechanisms to retain long-term temporal information, while GRUs simplify this structure, achieving similar performance at lower computational cost [52]. Transformers enable the modeling of long-range dependencies in parallel [54], while their extensions, GTN and TFT, further enhance this capability. GTN incorporates feature-wise attention to emphasize the most relevant variables [27], whereas TFT integrates static and dynamic inputs through gating mechanisms, making it the most advanced architecture in our study [35].

To ensure robust evaluation of the models, we implemented a 15-fold temporal cross-validation strategy spanning 2006–2022. In each fold, 14 years were used for training, one for validation, and the subsequent two consecutive years for testing. Validation always preceded the test period to reduce information leakage. Best-performing checkpoints were selected based on the validation performance and then evaluated on the held-out test folds.

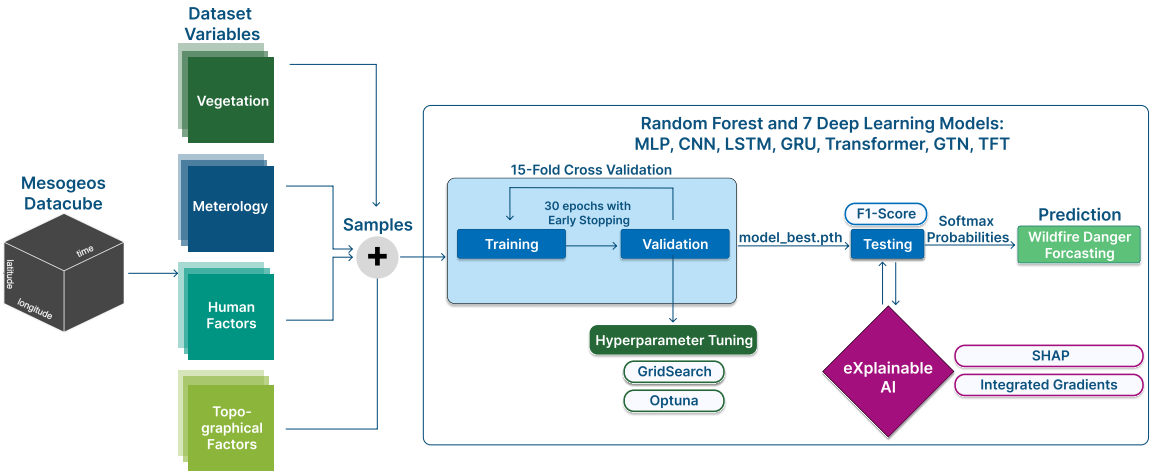

Figure 1: Overview of the machine learning (ML) pipeline. Samples extracted from the Mesogeos [27] data cube are used to train, validate, and test deep learning (DL) models via a cross-validation scheme. Final models are evaluated on the test set and interpreted using explainable AI (xAI) techniques.

2.3 Evaluation Metrics

We formulated wildfire prediction as a binary classification problem. Models output class probabilities via a Softmax layer, which are interpreted as the predicted level of fire danger. For evaluation, a fixed threshold of 0.5 is applied, which means that probabilities above this threshold are classified as fire (positive, high fire danger), and those below as no fire (negative, low fire danger).

To assess model performance, we use the F_1-score [57], which is well-suited for imbalanced classification where wildfire occurrences are rare compared to non-fire or negative events [58].

The F_1-score (Equation (1)) is the harmonic mean of precision and recall, balancing false positives and false negatives.

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

F_1 values range from 0 to 1, where 1 denotes perfect precision and recall, and values closer to 0 indicate indicate poor performance in either precision or recall or both metrics.

2.4 Explainable AI

Despite their impressive predictive power, many ML models, particularly deep neural networks, offer limited transparency in their decision-making processes. To address this limitation, and given the growing demand for interpretable methods in climate science [59], this study employs attribution-based xAI techniques to uncover how input features shape fire predictions. Because different explanation methods capture distinct theoretical notions of feature influence, combining approaches allows a more comprehensive view of model behavior and helps mitigate the weaknesses of any single approach [59]. In this work, we used mainly SHAP values [60], which provide a game-theoretic framework to assign each feature an importance score by estimating its marginal contribution across all possible feature coalitions. SHAP values should be understood as the estimated contribution of a given feature value to the difference between the model's prediction for the current instance and the mean model prediction, conditional on the current set of feature values [61]. In practice, we applied Kernel SHAP, a model-agnostic approximation method [29] implemented in the SHAP Python library [60], which balances accuracy and computational cost by requiring fewer model evaluations than other sampling-based approaches [62]. Because xAI techniques can yield varying results depending on model structure and data characteristics, we complemented our SHAP-based analyses with Integrated Gradients (IG), reported mainly in the Supplementary Information to compare and validate attribution patterns. IGs [63] quantify feature influence by computing gradients along a straight-line path from a baseline input x' to the actual input x, and to accumulate these gradients, thereby capturing how changes in each feature affect the prediction. In this work, IGs are implemented using the *Captum* library, a PyTorch-based framework [64] for model interpretability. The library provides a high-level interface for computing attributions across various neural network architectures without requiring changes to the

underlying model. Both methods aim to explain model behavior without altering the underlying architecture, with SHAP offering local instance-level explanations and IG providing path-averaged attributions that are only suited for deep neural networks. Other xAI approaches could include off-the-shelf RF feature importance [65, 24] or the intrinsic interpretability of attention weights in Transformers [66]. However, these methods are architecture-specific and therefore not suitable for our study, which spans heterogeneous model classes. For xAI analyses, we used the fixed chronological split from the original Mesogeos study (training: 2006–2019, validation: 2020, testing: 2021–2022), as this setup best reflects real-world deployment conditions where models are applied to unseen future data.

3 Results

3.1 Overall Model Performance

All ML models achieved strong classification performance, with mean F_1-scores above 0.75 on the test set (Figure 2). To formally test performance differences among models, we applied both parametric (ANOVA F-test) and non-parametric (Kruskal–Wallis [67]) analyses. Both tests rejected the null hypothesis of equal model performance ($p < 10^{-8}$). Post-hoc Dunn tests further confirmed that the RF baseline performed significantly worse than all neural network models ($p < 0.03$ for all pairwise comparisons), consistent with prior wildfire prediction studies [28, 23] in which DL models outperformed the RF baseline. However, subsequent analyses revealed that Transformer models maintained particularly stable and robust performance under varying temporal conditions (see Section 3.3).

Among the DL architectures, Transformer-based architectures achieved the highest and most consistent performance across test sets. However, no statistically significant differences were detected among the DL architectures (see Figure S5 for detailed results). Notably, the more complex attention-based variants, the TFT and GTN, did not exhibit significant improvements over the baseline Transformer model.

Compared to the Mesogeos Track A benchmarks [27], our re-implementation of baseline architectures (LSTM, Transformer, GTN) achieved slightly higher F_1-scores through cross-validation and extensive hyperparameter tuning, exceeding the previously reported performance of approximately 0.78. Incorporating additional architectures further increased overall performance, with nearly all DL models reaching or exceeding F_1scores values of 0.80.

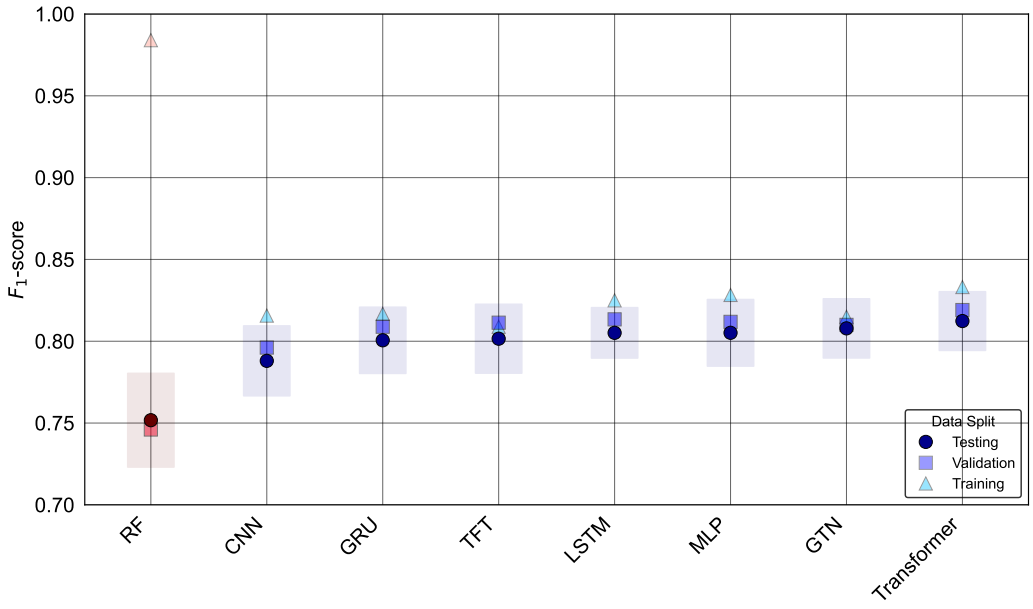

Figure 2: Cross-validated F_1-scores for all models. Mean values for training, validation, and testing across cross-validation splits are shown, with error bars indicating the standard deviation of the testing performance. Models are ordered by average test scores. The Random Forest model exhibits pronounced overfitting compared to the neural network architectures and is the only model with high test variability alongside low training variability, whereas the DL models show the opposite pattern, more spread in training and tighter, more stable performance on the test set.

3.2 Spatial Evaluation of Model Performance

To complement quantitative performance metrics, we visualized the spatial distribution of softmax probabilities predicted by the Transformer on the test set. In these maps, red points denote predicted fires (positives), while blue points indicate predicted non-fire events (negatives). As shown in Figure 3, clear spatial heterogeneity emerges, in which model performance is stronger in coastal regions (e.g., Croatia, Albania, Greece, and Sicily), while misclassifications occur more frequent inland. This pattern suggests that training data imbalances, such as the disproportionate number of fire events in coastal versus inland regions, may influence model performance, reflecting a common challenge in ML [68, 69, 70].

To test whether this pattern originates from dataset characteristics, we computed the Euclidean distance of each positive sample to the nearest coastline, defining fires within 100 km as coastal. Overall, 73% of positive fire events fall within this coastal zone, while 27% occur inland. While the overall class distribution in the dataset is already imbalanced with approximately two thirds of samples labeled as negative and one third as positive, this imbalance is further exacerbated geographically. If most of the positive samples are concentrated in coastal areas and the majority of negative samples are located inland, the model may struggle to generalize well to underrepresented inland fire events.

Beyond the overall performance metrics, we computed separate coastal and inland F_1-scores, precision, and recall on the test data to quantify the coastal–inland effect. Across all architectures, coastal performance exceeded inland performance (Figure S1). Among the DL models, coastal–inland differences were consistently pronounced (on average $\Delta F_1 \geq 0.17$), with the largest gap observed for the Transformer ($F_{1,\text{coastal}} = 0.84$ vs. $F_{1,\text{inland}} = 0.62$, $\Delta F_1 = 0.226$), whereas the RF exhibited only a modest difference ($\Delta F_1 = 0.05$).

Statistical testing across the eight models confirms that these performance differences are robust: Wilcoxon signed-rank tests, a suitable non-parametric alternative to the paired t-test given the small sample size [71], show that coastal predictions significantly outperform inland predictions in precision (0.80 vs. 0.69, $p = 0.0078$, $r = 0.89$), recall (0.88 vs. 0.74, $p = 0.0078$, $r = 0.89$), and F_1 score (0.84 vs. 0.71, $p = 0.0080$, $r = 0.89$).

To further examine whether this F_1-score disparity is merely a consequence of the smaller inland sample size, we tested a region-specific modeling strategy by subdividing the dataset and training separate models for coastal and inland areas (Table 1). This idea was motivated by Schmitt et al. [70], who subdivided California into three ecosystem-based prediction zones, the Southern and Northern California coasts and the Central Sierra Nevada mountain inland region, to investigate whether such regional separation could improve model performance under imbalanced data conditions. In our case, we used the full set of available inland fire samples and randomly selected an equal number of coastal samples from the training years to ensure comparable data volumes between models, keeping all hyperparameters constant as in the overall performance assessment above.

Overall, the mean inland F_1-score increased by $\Delta F_1 \approx 0.043$ across models relative to the one-model baseline, with the Transformer showing the largest gain ($\Delta F_1 = 0.1104$). This suggests that models trained on more homogeneous environmental conditions can better capture region-specific fire dynamics. By contrast, coastal performance decreased slightly when training two separate models, likely due to the intentionally smaller sample size. Nevertheless, our findings confirm that even in the two separate models case we can still see the systematic effects in reduced inland performance as the coastal–inland gap remains substantial in the two-model case (mean $\Delta F_1 \approx 0.068$), even under region-specific training. An exception was the Random Forest, whose inland F_1-score decreased by $\Delta F_1 = -0.2299$ when trained solely on inland samples, suggesting it benefits primarily from larger overall sample sizes regardless of spatial origin.

	One model		Two separate models (equal sample size)	
Model	F1–Coastal	F1–Inland	F1–Coastal	F1–Inland
CNN	0.82	0.63	0.78	0.71
GRU	0.82	0.63	0.80	0.74
GTN	0.84	0.66	0.77	0.69
LSTM	0.82	0.63	0.78	0.74
MLP	0.81	0.64	0.78	0.72
RF	0.91	0.86	0.73	0.63
TFT	0.83	0.66	0.78	0.71
Transformer	0.85	0.62	0.79	0.73

Table 1: F_1-scores for coastal and inland predictions under two training setups: a single unified model and two separate models trained with equal sample size. In both setups, coastal F1-scores exceed inland, though the average coastal–inland gap is smaller with two separate models.

3.3 Performance Dependence on Temporal Context Length

We further evaluated the impact of historical input length (5–30 days) on predictive performance (Figure 3b). Across all time horizons, the Transformer consistently outperformed the LSTM. Its performance increased nearly linearly with longer input sequences up to about 25 days, after which it began to plateau. This result demonstrates the advantage of attention-based models in capturing longer range temporal dependencies [54, 72]. In contrast, LSTM performance saturated after around 15 days and declined for longer input sequences, reflecting its limited ability to retain long-term information, likely due to the still remaining vanishing gradient problems, and the curse of dimensionality [66]. These results support prior studies [26, 73], which also reported that recurrent architectures tend to plateau with increasing temporal context, whereas Transformers can more effectively exploit longer input horizons.

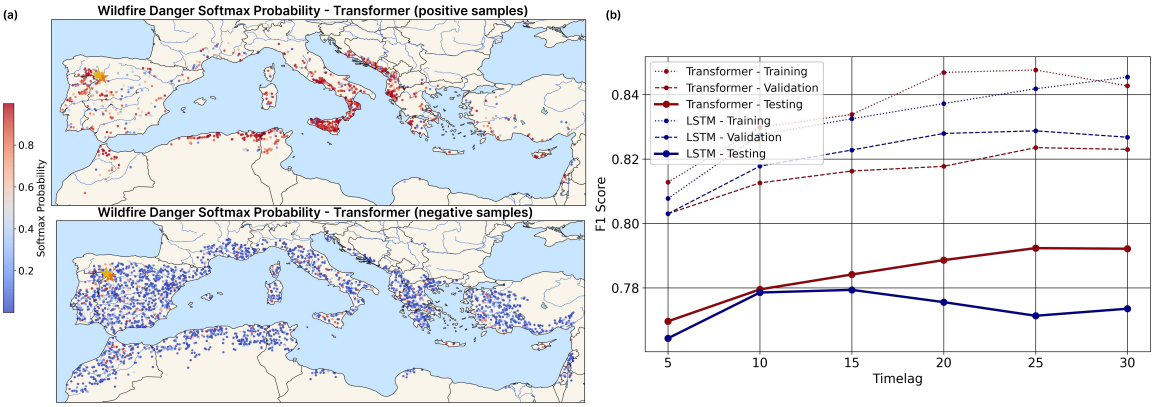

Figure 3: Spatial and Temporal Evaluation of Model Behavior. **(a)** Softmax probabilities predicted by the Transformer for positive (top) and negative (bottom) test samples, where red indicates high fire danger and blue indicates low fire danger. The two stars (in yellow and orange) mark the locations of two of the biggest fires in the test set, which are further analyzed in the case study in Section 3.6. **(b)** F_1-scores of the Transformer and LSTM models across different temporal input lengths (5–30 days) for training, validation, and testing. For each input length and model, a cross-validation was conducted with 10 different random initializations. In each case, the model achieving the highest validation F_1-score was selected and subsequently evaluated on the test set.

3.4 Explaining Models via SHAP Values

To assess the general relevance of input features for wildfire danger prediction, we computed mean absolute SHAP values for each feature, averaged across all ML models. Figure 4a) shows the features in descending order of importance, where higher values indicate a stronger influence on model predictions. Among all predictors, daytime land surface temperature *(lst_day)* emerges as the most important feature, followed by key meteorological drivers such as *relative humidity (rh)*, *2m temperature (t2m)*, and *total precipitation (tp)*. In contract, static land-cover variables consistently exhibited low SHAP values and small standard deviations, indicating that, for daily wildfire predictions in the Mediterranean region, variations in land cover contribute less to our

ML-based wildfire prediction than meteorological variations. This finding aligns with expectations, as daily fire risk is primarily driven by weather, whereas "suitable" land cover and vegetation states define the underlying conditions for fire occurrence. A corresponding analysis using IGs for the deep-learning models (Figure S2a) yielded broadly consistent results, likewise highlighting *lst_day* as the most influential predictor. However, IG showed larger variability across models (i.e., higher standard deviations of mean absolute attributions) and slightly lower importance for correlated temperature variables, an effect that may partly arise from the choice of a zero baseline in the IG computation [74]. Further investigation is needed to systematically assess how different baseline choices affect the stability and comparability of IG-based attributions.

Complementing the mean SHAP values, Figure 4b presents a heatmap of feature ranks across ML model types. Recurrent architectures such as LSTM and GRU exhibit nearly identical feature importance patterns, reflecting their similar structure, while RF differs markedly from DL models. RF emphasized variables such as soil moisture and slope, while down-weighting meteorological variables such as *lst_day* and *rh* that dominate in neural architectures. This discrepancy may arise from architectural differences, as tree-based models concentrate attribution on a few strong predictors via hierarchical splitting, whereas DL models distribute importance more evenly across interacting features. For comparison, the corresponding IG-based feature ranks are shown Figure S2b. The rank patterns are broadly consistent with those obtained from SHAP.

Several caveats should be noted when interpreting these results. xAI techniques such as SHAP values reflect the models' learned behavior towards factors driving predictions, and not necessarily the causal dynamics of wildfires, particularly because data-driven models are primarily aimed at maximizing predictive power, and not physical relationships or causality [75, 76, 77]. Furthermore, the interpretation of SHAP values must be taken with caution due to strong correlations among temperature-related variables (e.g., *t2m*–*lst_night*: $r = 0.67$, see Figure S4), which may cause attribution scores to be split or inconsistently distributed across redundant features [78, 79]. Moreover, we also note the intrinsic uncertainties in SHAP value approximation methods, especially in (close) relative variable rankings [80]. Despite these limitations, our analysis provides valuable insight into the models' decision-making processes and evaluates whether the learned behavior are consistent with established scientific understanding of wildfire drivers.

Finally, we highlight that maximizing predictive power is not the only reason for keeping the complete input variable set. For example, excluding variables such as *t2m* due to their strong correlation with other important temperature variables is not advisable, as they offer robust and gap-free coverage, unlike satellite-derived products that frequently contain missing values. They thus offer additional information in times of missing values to some, otherwise higher ranked, variables. In cases, such information might prove essential to realistically assess wildfire risk.

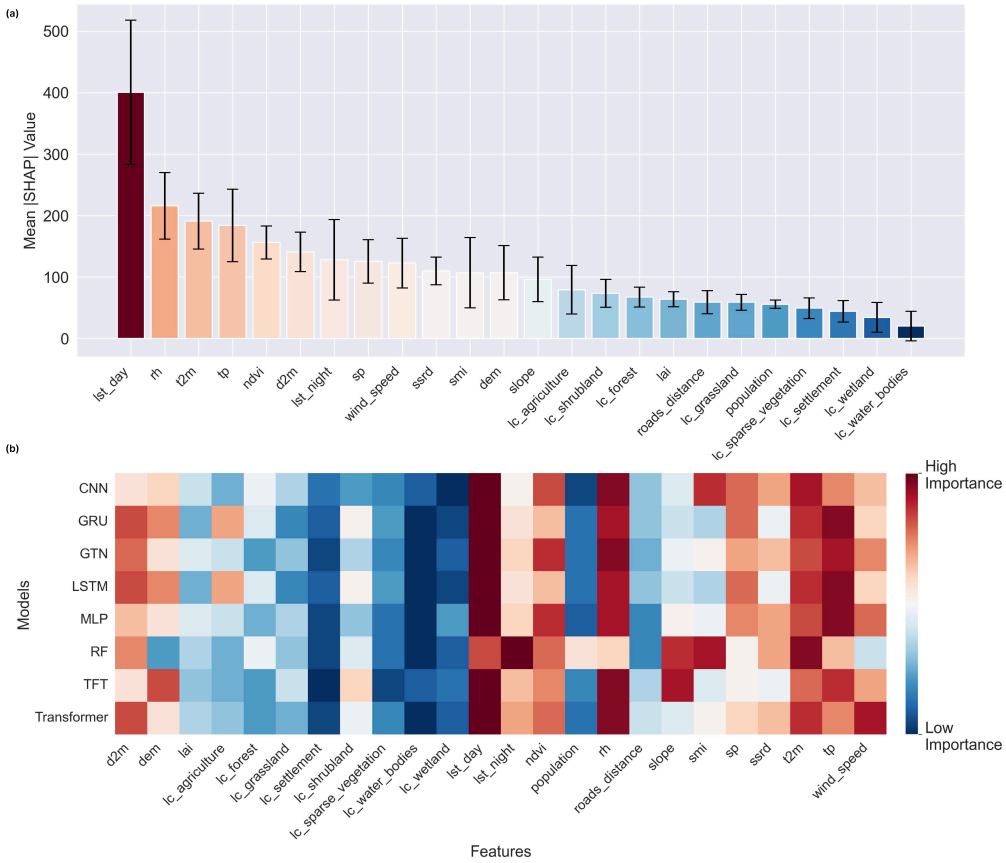

Figure 4: Cross-Model SHAP-based Feature Analysis. (a): average absolute SHAP values per feature across all models, sorted in descending order of mean importance. Higher values indicate stronger influence on predictions. Bars show the mean absolute SHAP value, and error bars denote the standard deviation across models. (b): heatmap of relative feature ranks across models, with dark red indicating higher importance and blue lower relevance.

3.5 Model Alignment with Physical Domain Knowledge

Building on the work of Li et al. [23], who emphasized the importance of evaluating models beyond predictive accuracy, we assessed the physical consistency of our ML models. For each feature, we compared the sign of its SHAP value with the normalized input value relative to the expected direction of effect derived from fire behavior literature (Table S2). A sample was considered physically consistent if, for positively related features, high input values were associated with positive SHAP values (or low inputs with negative SHAP values), and analogously for negatively related features. The proportion of physically consistent samples per feature defines the physical consistence score, illustrated in Figure 5. For example, there are clear positive relationships for temperature variables like *t2m*, *d2m*, *lst_day*, and *lst_night* that promote fuel drying and thus fire ignition, aligned with findings by Chuvieco et al. [81] and Di Giuseppe et al. [21]. Only 19 out of the 24 predictors were included in this analysis, excluding variables that exhibit ambiguous or highly context-dependent relationships with fire occurrence [82, 83], such as *ndvi*, *population*, *elevation (dem)*, *surface pressure (sp)*, and *leaf area index (lai)*.

Among the DL architectures, the Transformer and GTN models achieved the highest degree of physical consistency, correctly capturing 11 of 19 relationships. In contrast, simpler models such as MLPs or LSTMs captured fewer consistent associations. Interestingly, the RF baseline model with the lowest F_1-scores outperformed all others in terms of physical consistency, capturing 13 out of 19 relationships, including several land cover classes representing different fuel availabilities (e.g., *lc_settlement*, *lc_waterbodies)* that were not correctly represented by any DL model. Certain variables, mainly static predictors such as *lc_wetland*, *lc_shrubland*, and *wind_speed*, were not correctly captured by any model, indicating limited relevance for data-driven fire occurrence prediction. Future work might explore if results would differ when other aspects of wildfire events are predicted, such as the spatial extent of the associated burnt area, which might be much more dependent on, e.g., wind speed than mere fire occurrence. In addition, it is unclear if the data-driven models can cleanly separate, e.g., wet-windy from hot-windy days, which would

naturally be subject to very different wildfire risk, covering a continuum of weather states.

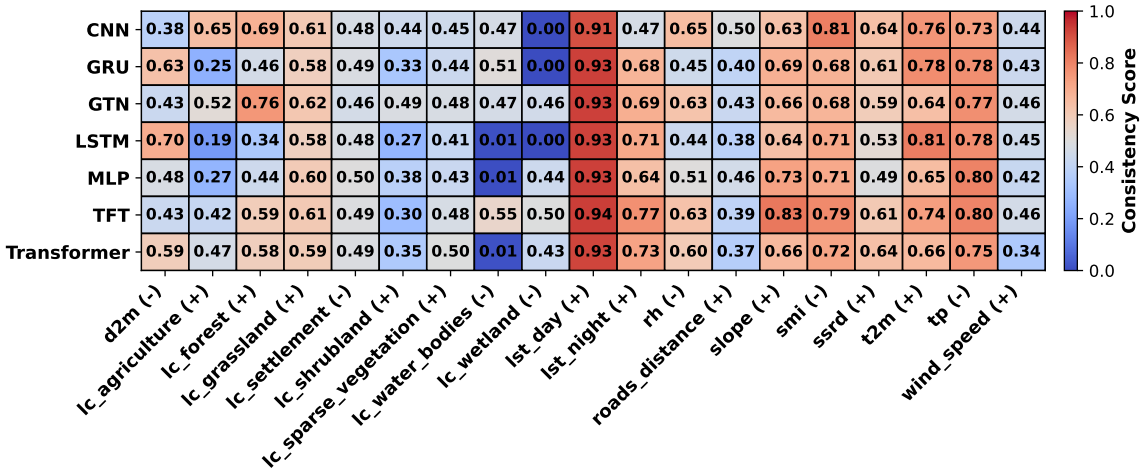

Figure 5: Physical consistency scores showing, for each feature–model pair, the proportion of samples in which SHAP-based attributions align with expected physical relationships from the fire science literature. The sign of the expected relationship with wildfire occurrence is indicated for each variable by ± in brackets. Darker colors highlight higher consistency within a model; the raw values are also provided as numbers.

3.6 Case Studies: Comparison of Two Fire Events in Spain

Our xAI analyses so far focused on overall model decision process. However, they could not reveal how the data-driven models arrived at individual predictions or how sensitive these predictions are to specific input features, including the potential impact of collinear variables. To address these questions and better understand the trade-offs between accuracy and explainability, we examined two large fire events in Spain's Zamora province (Castilla y León) during summer 2022. We analysed a false negative on June 15 and a true positive on July 17 (Figure S6). These case studies are indicated by star symbols in Figure 3 (yellow for June, orange for July). Despite their similar magnitudes and close geographic proximity, the Transformer model assigned a low probability to the June event but correctly detected/predicted the July fire.

Figure 6 summarizes the temporal evolution of environmental conditions prior to each event. The July fire was preceded by a distinct intensification of fire-conducive conditions, including steadily rising surface and air temperatures (*t2m*, *lst_day*, *lst_night*) and decreasing soil moisture (*smi*), relative humidity (*rh*), and vegetation indices (*lai*, *ndvi*). These signals were less pronounced in the June case, which was reportedly ignited by lightning [84].

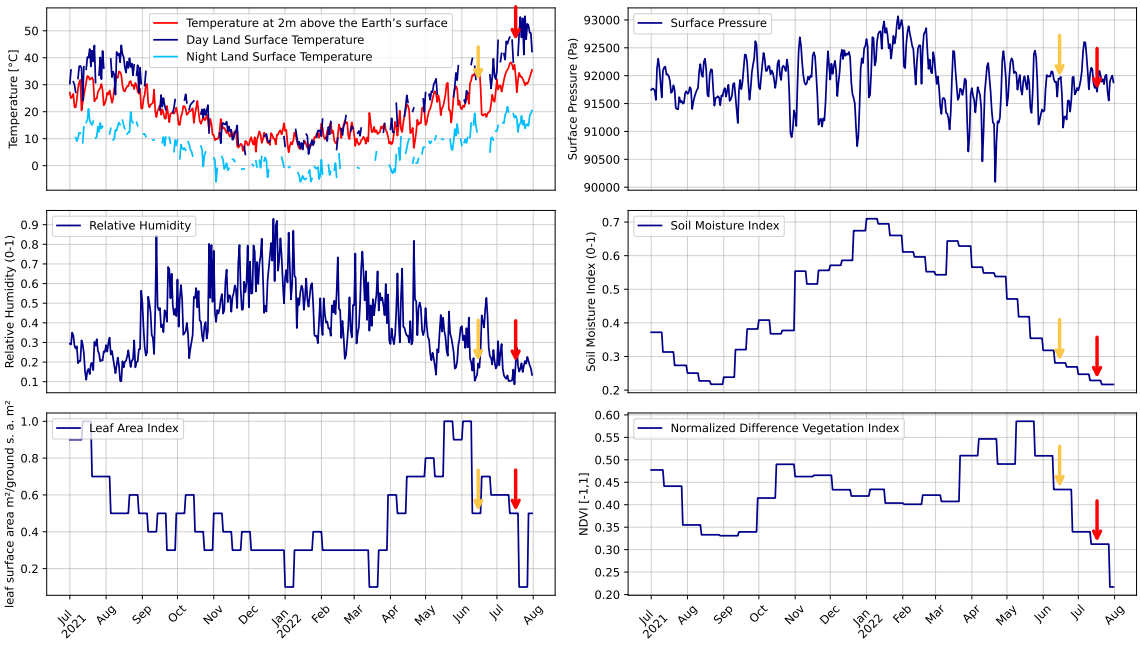

Figure 6: Temporal evolution of environmental variables before the two Spain fires 2022. The two arrows indicate the ignition dates: yellow for the June fire and red for the July fire. The data is shown at the average location between the two ignition points (41.82°N, 6.13°W). The two ignition points are only about 23.4 km apart in a straight line and are located at similar elevations. The plots show a sharp rise in air and surface temperatures during the day accompanied by a decrease in soil moisture and relative humidity. Note that wildfire danger predictions are based on a 30-day temporal context window prior to each event, rather than instantaneous values.

In Figure 7a, SHAP attributions explain the contrasting predictions. while the June event exhibits a strongly negative total SHAP sum suppressing fire probability, the July event shows strong positive contributions, mainly from temperature and humidity variables. Lower relative humidity in July (*rh*) is particularly decisive, driving high positive SHAP values, whereas higher humidity in June contributes less positively. Consistently, a higher soil moisture index (*smi*) in June is associated with strong negative SHAP values, reinforcing the suppressing effect of humid conditions. This highlights how meteorological conditions during a heatwave and drought shifted feature contributions to favor a fire prediction in July but not in June.

To test model reliance on land surface temperature, we performed an ablation experiment by retraining the Transformer without the *lst_day* feature as shown in Figure 7b. For the June fire, the predicted probability increased from 0.19 to 0.52, crossing the decision threshold and leading to a correct classification. Excluding *lst_day* also redistributed attributions, with other temperature-related variables (*t2m*, *d2m*) receiving stronger positive influence. This suggests that *lst_day* can suppress predictions under certain conditions, although the effect may depend on interactions with correlated drivers. DL architectures maintained high performance even when key variables were removed. Broader analyses across more events, however, would be required to determine whether such attribution instabilities reflect systematic mechanisms behind misclassifications.

To further validate these findings, we computed the IG attributions for the same two fire events (Figure S3). The Integrated Gradients were computed using a zero baseline (i.e., all input features set to zero), representing an absence of signal from which the contribution of each feature was accumulated. Unlike the SHAP results, where the June event exhibited a strongly negative total attribution, the IG analysis yielded positive total attributions for both fires, though markedly higher for the July case. The stronger total IG value for July again reflects higher model confidence during the extreme heat and drought conditions, whereas the June case exhibits weaker positive attributions, with high soil moisture and relative humidity reducing the overall fire likelihood. When the *lst_day* feature was excluded, the total IG attribution further increased, consistent with the SHAP-based ablation results.

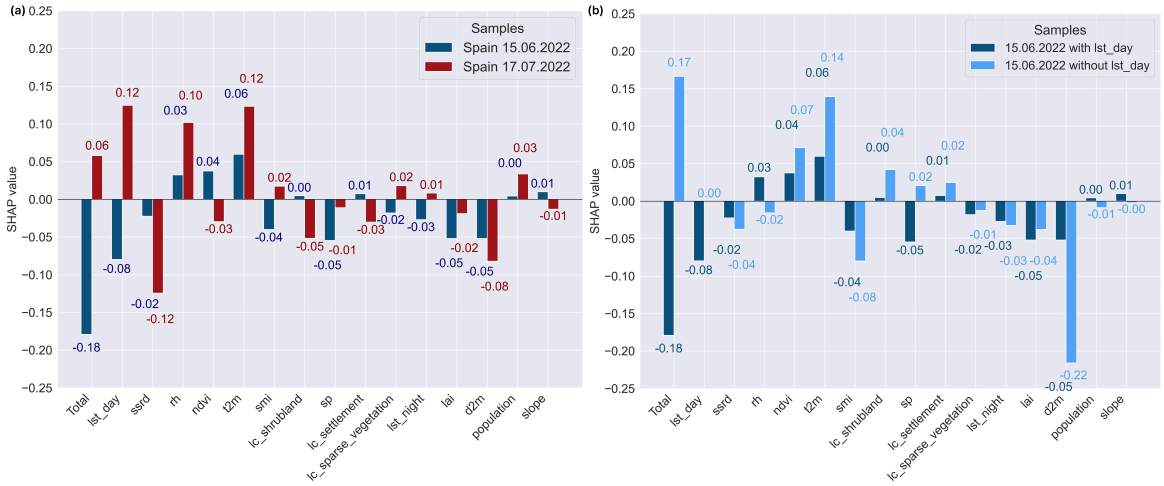

Figure 7: **(a)** SHAP attributions for the June fire (false negative in blue) and the July fire (true positive in red), highlighting differences in the influence of temperature- and humidity-related drivers. **(b)** SHAP attributions for the June fire with (dark blue) and without (light blue) the *lst_day* feature, showing that its exclusion increases fire probability and shifts contributions to other temperature variables. The absolute SHAP values are displayed above the bars.

4 Discussion & Conclusion

We provided a comprehensive and comparative evaluation of a diverse set of seven temporal DL architectures for daily wildfire danger forecasting in the Mediterranean, benchmarked against a RF baseline and complemented with xAI analyses to examine the trade-offs between predictive performance and model plausibility.

All DL models achieved good predictive performance ($F_1 > 0.75$), outperforming the RF baseline ($p < 0.01$), in agreement with Kondylatos et al. [28], who similarly reported superior DL performance over RF for wildfire danger forecasting in Greece. Performance differences among DL models were not statistically significant at the 95% confidence level. Notably, complex attention-based variants such as TFT and GTN did not surpass a standard Transformer, suggesting that simpler attention mechanisms may be sufficient to capture relevant temporal dependencies. Given the need for models that are not only accurate but also practical and trustworthy in operational fire danger forecasting, the Transformer emerges as the most suitable choice for real-world deployment. This finding aligns with prior work. Kondylatos et al. [28] reported that a "vanilla" LSTM surpassed the more complex ConvLSTM, and Di Giuseppe et al. [21] concluded that greater ML complexity does not inherently improve forecasts. This consistency reinforces that the quality of the data and its representation of the fire triangle (i.e., weather, fuel and ignition) is often a more critical factor than architectural sophistication [21]. Nevertheless, our temporal analysis confirmed that Transformers effectively exploit extended historical contexts more efficiently than LSTMs, with performance improving up to 30-day input windows. The LSTM plateaued earlier, reflecting its limitation in capturing longer temporal dependencies. This corroborates with Prapas et al. [26], who conducted a similar time lag sensitivity analysis using the global SeasFire data cube [85], evaluating forecasting windows up to 128 days with 8-day temporal resolution. While not directly comparable due to differences in temporal granularity and prediction targets, their findings similarly indicate that Transformer-based models degrade more gradually and plateau later with increasing predictor dimensionality. Similarly, Michail et al. [73] showed that models trained with longer time series achieve better and more stable performance but eventually saturate when attempting very long-range forecasting, particularly for recurrent architectures such as GRU or LSTM.

Our xAI analyses provided further critical insights, revealing a fundamental trade-off between model accuracy and plausibility. While the Transformer achieved the highest predictive accuracy, the RF benchmark captured overall more physically consistent relationships, including a few that no DL model represented correctly. This raises an important question: how can we define the "ideal" wildfire forecasting model? Should preference be given to models such as Transformers with superior predictive accuracy, or to models such as RF that here exhibit stronger physical plausibility (in terms of the overall tendency of predictor-predictand relationships) but apparently at the expense of lower predictive skill? Lower predictive skill could lead to dangerous false classifications, with potentially catastrophic consequences. Pure performance optimization might

make overall better risk classifications on standard testing data, but seemingly often for the wrong reasons. The answer certainly depends on the intended application, whether the goal is operational forecasting or scientific understanding. In the latter case, less robust physical predictor-predictand relationships could prove critical, e.g. when studying possible drivers of changes in wildfire risks under climate change scenarios [86, 87, 88]. The xAI analysis also showed that different DL model architectures broadly agreed on the most influential features (Figure 4b). Notably, both SHAP and IG revealed mostly consistent patterns in feature importance, reinforcing the robustness of the identified drivers across explainability methods. The three most relevant predictors were temperature, relative humidity, and precipitation, followed by NDVI and d2m. Although these results align with intuitive expectations and prior studies [28, 29], caution is warranted to avoid confirmation biases [89]. However, a critical mismatch with established wildfire dynamics was identified, the models failed to adequately capture the well-documented influence of human activity. It is well-established that humans are responsible for most ignitions [90], using fire for agriculture and deforestation [91], while also suppressing it through land management [92, 93, 94]. This model shortcoming is likely due to the poor representation of human drivers in the input data, where they are often included only as coarse, indirect proxies (e.g., annual population density, static distance to roads). Furthermore, the complex, non-binary effects of these indirect variables—where the same factor can either increase or decrease risk depending on context—make them incompatible with standard physical consistency scoring, which typically assumes simpler, monotonic relationships.

We also note that training data imbalances (e.g., the proportion of coastal vs. inland fire events) lead to reduced predictive performance in inland areas, highlighting a common challenge in ML [68, 69, 70]. Adopting a region-specific modeling strategy by subdividing the dataset by region, and train separate models for coastal and inland areas, as demonstrated in Schmitt et al. [70] lead to improvements in the F_1-score of Inland fires by $\Delta 0.04$ by tailoring the model to more homogeneous conditions, it comes with the trade-off of reduced generalizability.

The key contribution of this study is its integrated framework, which combines state-of-the-art temporal DL architectures with explainability methods to evaluate daily wildfire danger forecasting. By jointly assessing predictive skill, explainability, and physical consistency, we demonstrate that DL models can deliver substantial gains in forecast accuracy while capturing physically meaningful drivers. However, our in-depth analysis - combining xAI, qualitative inspection of predictions, and targeted case studies - revealed that data quality and representation are paramount for building trustworthy models, a factor often obscured by aggregate performance scores. Specifically, we identified three primary data-centric challenges. First, a visual inspection of model outputs revealed geographic imbalances in predictive performance that were not captured by overall accuracy metrics. Second, xAI techniques indicated that well-documented fire drivers, such as human-related variables, were not consistently influencing model predictions. Finally, case studies of Spanish fires pointed to a fire misclassification due to the absence of key ignition factors in the input data. For instance, a known lightning-caused fire may have been undetectable by our models because of the lack of the necessary atmospheric data, a finding that was only possible through the incorporation of domain knowledge. We suggest that future work should place explainability and in-depth analysis at the forefront, exploring global datasets for improved generalization and hybrid or physics-informed architectures to better integrate fire-drivers interactions, also subject to more sophisticated representations of human influences on the land surface. Such approaches are crucial to bridge the gap between predictive performance and physical consistency, ultimately providing more robust predictions and deepening our scientific understanding of fire-driver interactions.

Acknowledgments
Support for this research was provided by the Karlsruhe Institute of Technology (KIT), in particular by the Chair for AI in Climate and Environmental Sciences. The authors gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Education and Research and the Ministry of Science, Research and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding program (https://www.nhr-verein.de/en/our-partners). HoreKa is partly funded by the German Research Foundation (DFG).

Funding
PN was partially funded by the UK Natural Environment Research Council (NERC), grant number NE/V012045/1. CN was funded by the Karlsruhe Institute of Technology, particularly through the Young Investigator Group Preparation Programme.

Author contributions
PB prepared the data, led the ML training and data analysis, and wrote the initial paper draft, supervised by PN and CN. All authors provided feedback on the initial draft and made substantial contributions to its revisions. PN suggested the study, in discussion with CN.

Data availability
The Mesogeos datacube and all datasets used in this study are openly accessible, with updated download links provided on the project website: https://orion-ai-lab.github.io/mesogeos/.The primary scripts for model training and analysis are available in a publicly accessible repository, accompanied by documentation to support result replication:
https://github.com/paulinebecker2002/mesogeos

Supplementary data

Abbreviation	**Description**
	Vegetation
lai	Leaf Area Index (leaf surface area m^2/ground surface area m^2)
ndvi	Normalized Difference Vegetation Index [-1;1]
lc_agriculture	Land cover class: agriculture [0;1]
lc_forest	Land cover class: forest [0;1]
lc_grassland	Land cover class: grassland [0;1]
lc_shrubland	Land cover class: shrubland [0;1]
lc_sparse_vegetation	Land cover class: sparse vegetation [0;1]
lc_settlement	Land cover class: settlement [0;1]
lc_water_bodies	Land cover class: water bodies [0;1]
lc_wetland	Land cover class: wetland [0;1]
	Meterology
d2m	Dewpoint temperature at 2m (K)
lst_day	Day land surface temperature (K)
lst_night	Night land surface temperature (K)
rh	Relative Humidity [0;1]
smi	Soil Moisture Index [0;1]
sp	Surface Pressure (Pa)
ssrd	Surface Solar Radiation Downwards (J/m^2)
t2m	Temperature at 2m above the Earth's surface (K)
tp	Total Precipitation (m)
wind_speed	Wind Speed (m/s)
	Human Factors
population	Population ($people/km^2$)
roads_distance	Distance from Roads (m)
	Topographical Data
dem	Elevation (m)
slope	Slope of the Area (rad)

Table S1: Categories of Wildfire Contributing Factors. Overview of all explanatory variables used in the fire prediction models, grouped into four categories: Fuel, Meteorology, Human Factors and Topographical Data.

Variable	Sign	References
d2m	–	[81]
lc_agriculture	+	[90]
lc_forest	+	[81, 95]
lc_grassland	+	[96, 97]
lc_settlement	–	[95]
lc_shrubland	+	[96, 81, 97]
lc_sparse_vegetation	+	[81]
lc_water_bodies	–	[95]
lc_wetland	–	[81]
lst_day	+	[81, 21]
lst_night	+	[21]
rh	–	[94, 98]
roads_distance	+	[90, 95]
slope	+	[99, 100]
smi	–	[94, 81]
ssrd	+	[81, 21]
tp	–	[94, 95]
t2m	+	[81, 21]
wind_speed	+	[99, 100]

Table S2: Physical Relationships between Wildfire Drivers and Fire Occurrence. The table summarizes the expected direction of influence (positive or negative) of key environmental and anthropogenic variables on wildfire occurrence, based on findings from recent literature. A positive sign (+) indicates a promoting effect on fire occurrence, while a negative sign (–) suggests a suppressing effect.

Model	**Hyperparameter Settings and Implementation Details**
Random Forest (RF)	Framework: scikit-learn – RandomForestClassifier; $n_{\text{estimators}} = 922$, max_depth = 32, max_features = `sqrt`, min_samples_split = 3, min_samples_leaf = 3, class_weight = balanced, random_state = 12345;

Multilayer Perceptron (MLP)	*Model size:* hidden dimensions [256, 128];
	Regularization: dropout 0.28;
	Optimizer/Training: Adam (learning rate 5.18×10^{-4}, weight decay 3.22×10^{-4});
	batch size = 128;
	epochs = 30;
	StepLR (step_size = 15, $\gamma = 0.1$);

Convolutional Neural Network (CNN)	*Input:* input channels = 1;
	Model size: conv feature dim = 128;
	fully connected head [256, 128];
	Regularization: dropout 0.021;
	Optimizer/Training: Adam (learning rate 1.45×10^{-3}, weight decay 1.23×10^{-3});
	batch size = 512;
	epochs = 30;
	StepLR (step_size = 15, $\gamma = 0.666$);

Long Short-Term Memory (LSTM)	*Model size:* SimpleLSTM (1 layer), hidden size = 128;
	Regularization: dropout 0.25;
	Optimizer/Training: Adam (learning rate 1×10^{-3}, weight decay 6.3×10^{-3});
	batch size = 256;
	epochs = 30;
	StepLR (step_size = 15, $\gamma = 0.1$);

Gated Recurrent Unit (GRU)	*Model size:* SimpleGRU (1 layer), hidden size = 128;
	Regularization: dropout 0.10;
	Optimizer/Training: Adam (learning rate 1.25×10^{-3}, weight decay 6.3×10^{-3});
	batch size = 128;
	epochs = 30;
	StepLR (step_size = 15, $\gamma = 0.1$);

Transformer	*Model size:* model dimension = 256;
	feed-forward dim = 512;
	heads = 2; *Input:* —
	layers = 2;
	hidden dimensions [128, 64];
	Regularization: dropout 0.34;

Gated Transformer Network (GTN)	*Model size:* TransformerNet with feature-wise (gated) channel attention;
	model dimension = 256;
	feed-forward dim = 512;
	heads = 4;
	layers = 4;
	Regularization: dropout 0.30;
	Optimizer/Training: Adam (learning rate 1.0×10^{-4}, weight decay 4.5×10^{-3});
	batch size = 128;
	epochs = 30;
	StepLR (step_size = 15, $\gamma = 0.1$);

Temporal Fusion Transformer (TFT)	*Input:* dynamic = 12, static = 12 (handled separately); sequence length = 30;
	Model size: TFTNet with gating and attention;
	model dimension = 128;
	heads = 8;
	layers = 2;
	Regularization: dropout 0.398;
	Optimizer/Training: Adam (learning rate 1.11×10^{-3}, weight decay 5.0×10^{-5});
	batch size = 256;
	epochs = 30; StepLR (step_size = 15, $\gamma = 0.397$);
	early stopping = 10;

Table S3: Optimized hyperparameters and implementation details for all models after cross-validation. Each model's configuration reflects the best-performing Optuna trial based on validation performance. All models were optimized within the same predefined hyperparameter ranges: hidden_dims $\in \{[128, 64], [256, 128], [512, 256]\}$, batch_size $\in \{128, 256, 512, 1024\}$, learning rate $\in [1 \times 10^{-5}, 1 \times 10^{-2}]$, dropout $\in [0, 0.7]$, weight decay $\in [1 \times 10^{-6}, 1 \times 10^{-2}]$, and scheduler decay factor $\gamma \in [0.1, 0.9]$. All deep learning models employ a two-unit softmax output with negative log-likelihood (NLL) loss, are implemented in PyTorch, and log training and validation with TensorboardWriter.

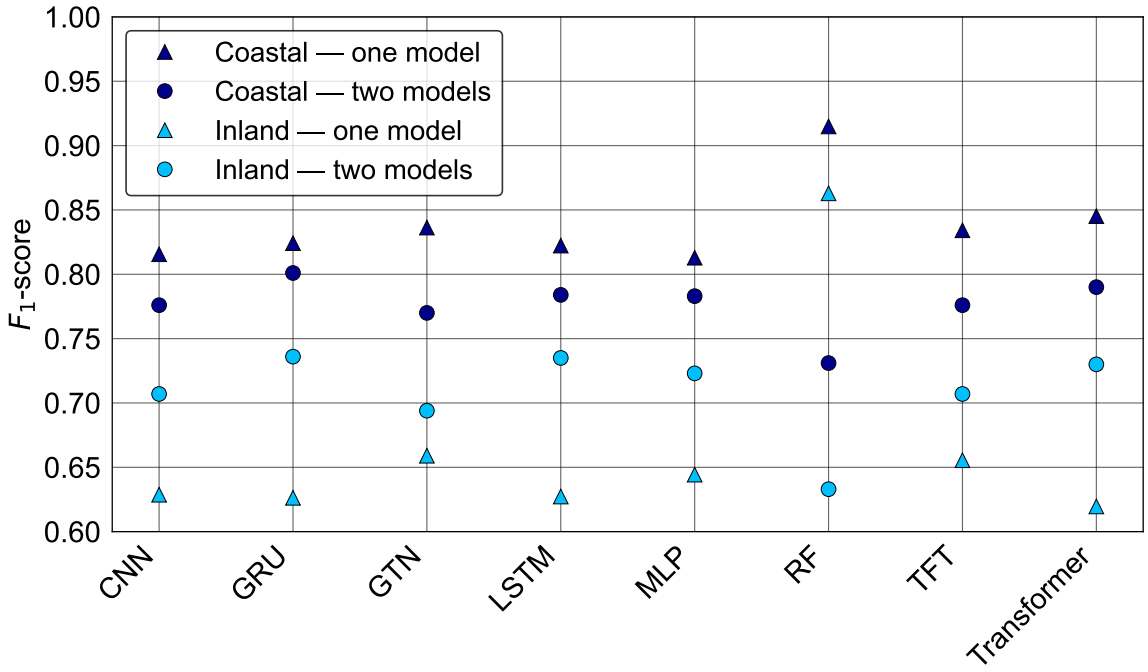

Figure S1: F_1-scores by model for coastal (triangles) and inland (circles) under two training setups: a single unified model (lightblue) and two separate models trained with equal sample sizes (dark blue). F_1-scores were calculated for the test set using a chronological split covering the years 2006–2019 for training, 2020 for validation, and 2021–2022 for testing. In both setups, coastal F_1-scores exceed inland, though the average coastal–inland gap is smaller with two separate models.

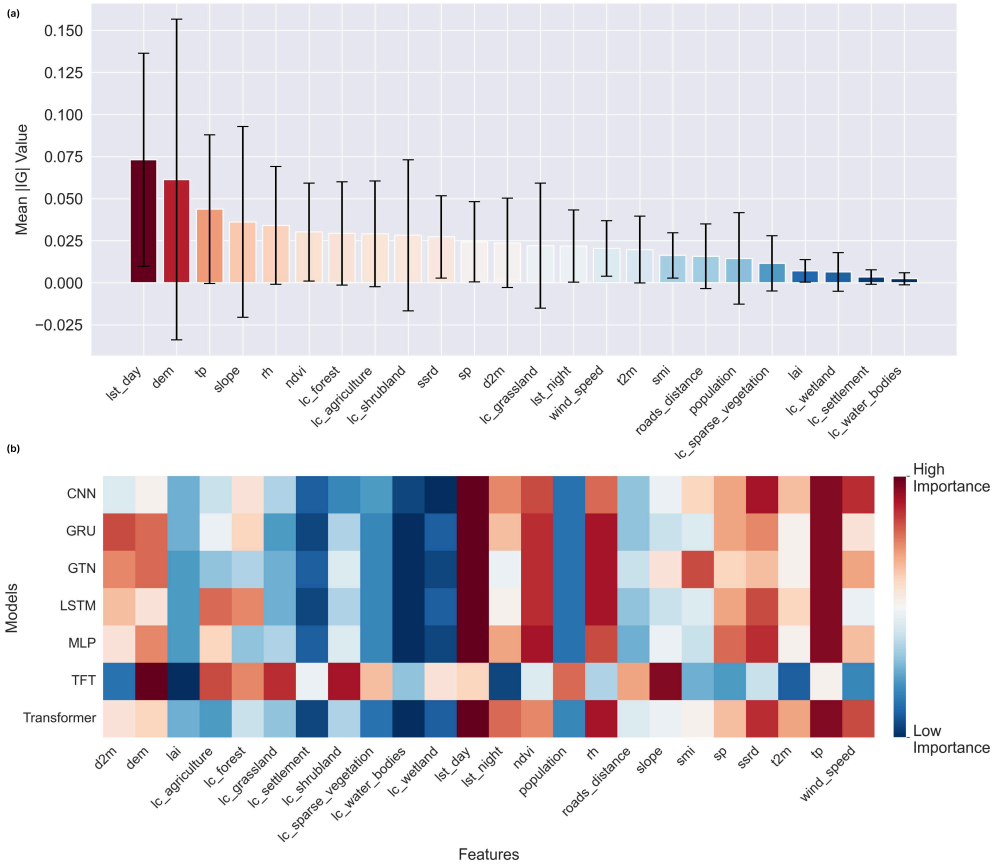

Figure S2: Cross-Model IG-based Feature Analysis. (a): Average absolute IG values per feature across all models, sorted in descending order of mean importance. Higher values indicate stronger influence on predictions. Bars show the mean absolute IG value, and error bars denote the standard deviation across models. (b): Heatmap of feature importance ranks derived from absolute IG values. Darker red tones indicate features consistently ranked as more influential across models, while blue tones represent features of lower relative influence. Note that the RF model is not included, as gradient-based attribution methods such as IG cannot be applied to non-differentiable, tree-based architectures.

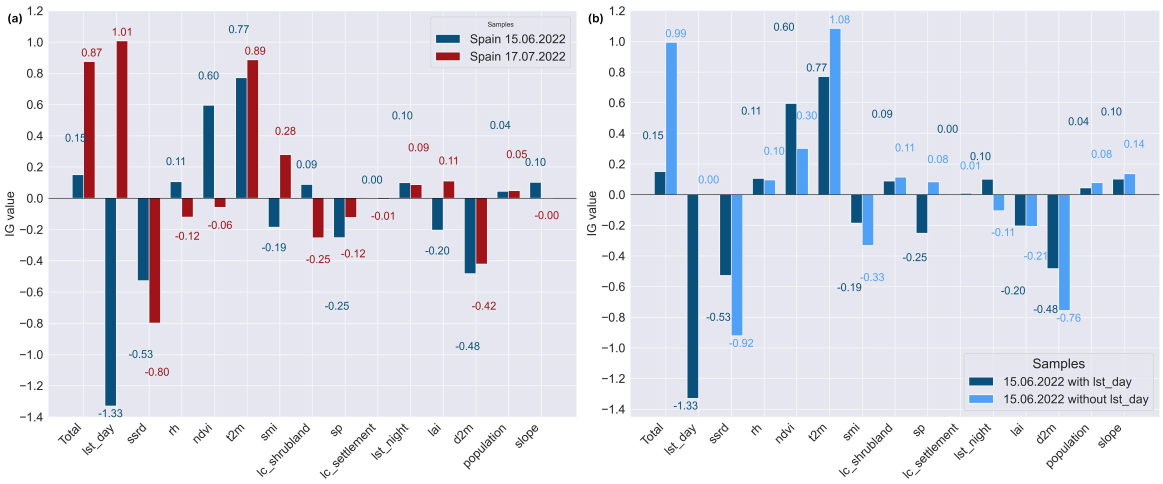

Figure S3: **(a)** IG attributions for the June fire (false negative) and the July fire (true positive), showing each feature's contribution to the final prediction. Both events exhibit a positive total IG attribution, indicating an overall tendency toward a fire prediction, but the July event shows a markedly higher total IG value, reflecting stronger model confidence driven by temperature- and humidity-related factors. **(b)** IG attributions for the June fire with and without *lst_day* showing that its exclusion increases the total positive attribution and thus strengthens the predicted fire probability. For comparability with Fig. 7, features are displayed in the same order, sorted by the descending absolute SHAP difference.

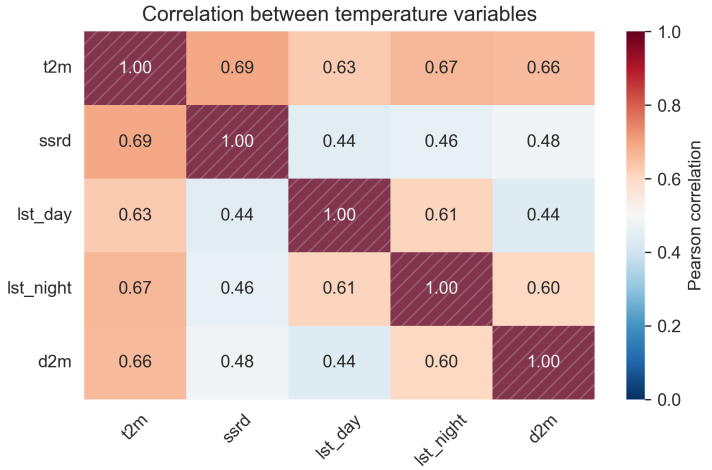

Figure S4: Correlation between Temperature-related Variables. High correlation coefficients highlight the redundancy between different temperature-based predictors. Diagonal cells are shaded to indicate perfect self-correlation.

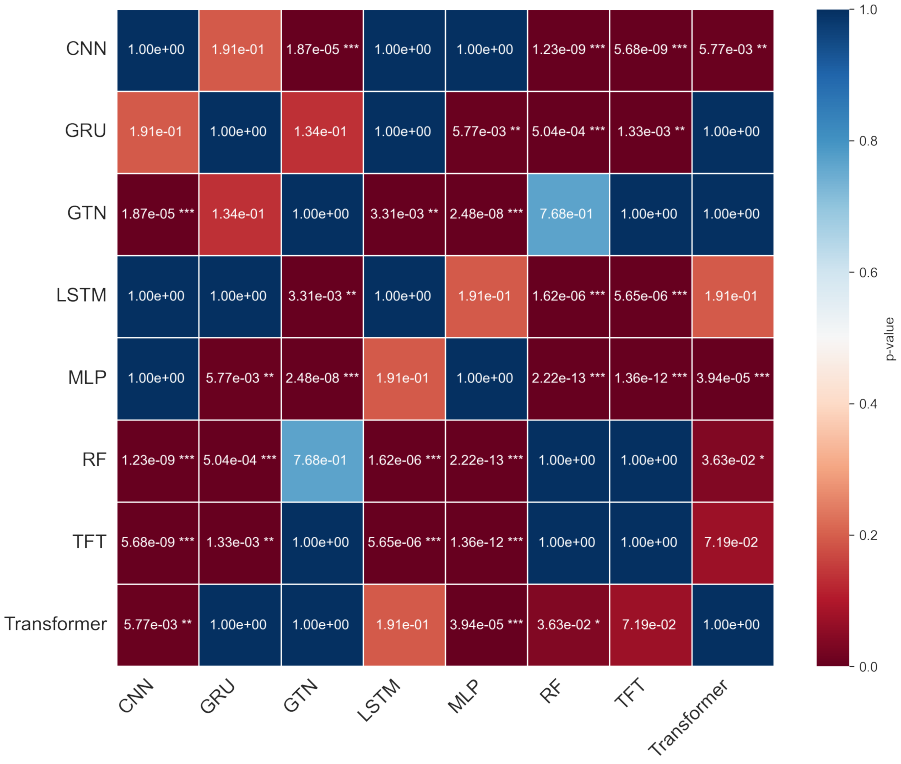

Figure S5: Post-hoc Dunn Test for Model Comparisons. Pairwise comparisons of model performance on the test set using Dunn's test with Holm correction. The heatmap shows adjusted p-values, where darker red cells indicate stronger evidence against the null hypothesis. Asterisks denote significance levels (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Results confirm that the RF baseline performs significantly worse than all DL models, while no significant differences were found among the DL architectures.

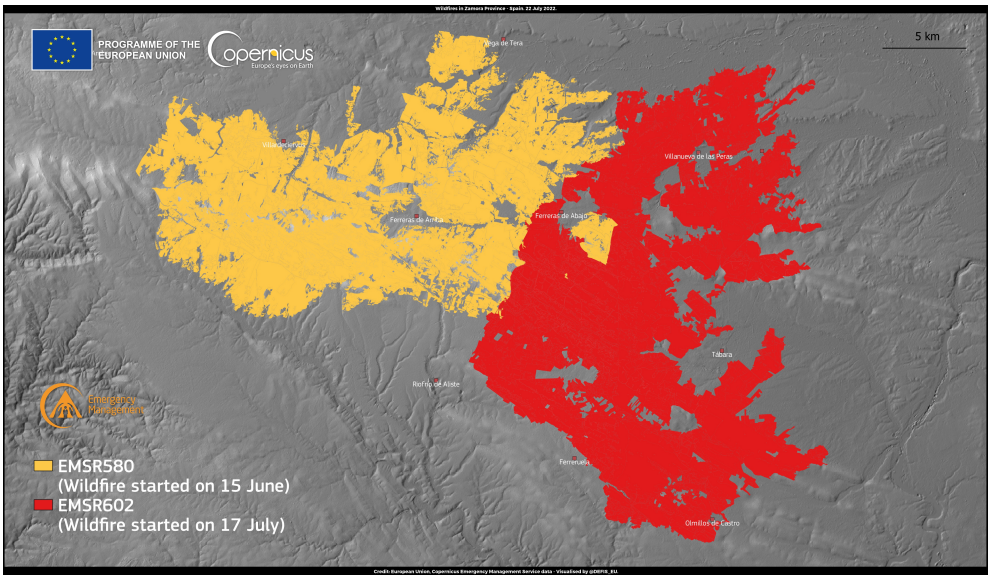

Figure S6: Burned Areas of Two Big Wildfires in Zamora Province, Spain 2022: EMSR580 started on 15 June (yellow), EMSR602 on 17 July 2022 (red), Credit: European Union, Copernicus Emergency Management Service Data [101]

References

1. Prapas I, Ahuja A, Kondylatos S, Karasante I, Panagiotou E, Alonso L, Davalas C, Michail D, Carvalhais N, and Papoutsis I. Deep Learning for Global Wildfire Forecasting.

2023. DOI: 10.48550/arXiv.2211.00534. Available from: http://arxiv.org/abs/2211.00534 [Accessed on: 2025 Apr 18]

2. Jones MW, Abatzoglou JT, Veraverbeke S, Andela N, Lasslop G, Forkel M, Smith AJP, Burton C, Betts RA, Werf GR van der, Sitch S, Canadell JG, Santín C, Kolden C, Doerr SH, and Le Quéré C. Global and Regional Trends and Drivers of Fire Under Climate Change. en. Reviews of Geophysics 2022; 60:e2020RG000726. DOI: 10.1029/2020RG000726. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1029/2020RG000726 [Accessed on: 2025 Jul 30]
3. Perkins O, Kasoar M, Voulgarakis A, Edwards T, Haas O, and Millington JDA. The Spatial Distribution and Temporal Drivers of Changing Global Fire Regimes: A Coupled Socio-Ecological Modeling Approach. en. Earth's Future 2025; 13. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2024EF004770:e2024EF004770. DOI: 10.1029/2024EF004770. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1029/2024EF004770 [Accessed on: 2025 Oct 24]
4. Roșu IA, Mourgela RN, Kasoar M, Boleti E, Parrington M, and Voulgarakis A. Large-scale impacts of the 2023 Canadian wildfires on the Northern Hemisphere atmosphere. en. npj Clean Air 2025 Sep; 1. Publisher: Nature Publishing Group:22. DOI: 10.1038/s44407-025-00022-9. Available from: https://www.nature.com/articles/s44407-025-00022-9 [Accessed on: 2025 Oct 24]
5. Grillakis M, Voulgarakis A, Rovithakis A, Seiradakis KD, Koutroulis A, Field RD, Kasoar M, Papadopoulos A, and Lazaridis M. Climate drivers of global wildfire burned area. en. Environmental Research Letters 2022 Apr; 17. Publisher: IOP Publishing:045021. DOI: 10.1088/1748-9326/ac5fa1. Available from: https://doi.org/10.1088/1748-9326/ac5fa1 [Accessed on: 2025 Oct 24]
6. Zheng B, Ciais P, Chevallier F, Yang H, Canadell JG, Chen Y, Velde IR van der, Aben I, Chuvieco E, Davis SJ, Deeter M, Hong C, Kong Y, Li H, Li H, Lin X, He K, and Zhang Q. Record-high CO2 emissions from boreal fires in 2021. Science 2023 Mar; 379. Publisher: American Association for the Advancement of Science:912–7. DOI: 10.1126/science.ade0805. Available from: https://www.science.org/doi/10.1126/science.ade0805 [Accessed on: 2025 Oct 24]
7. Migliavacca M, Dosio A, Camia A, Hobourg R, Houston-Durrant T, Kaiser JW, Khabarov N, Krasovskii AA, Marcolla B, San Miguel-Ayanz J, Ward DS, and Cescatti A. Modeling biomass burning and related carbon emissions during the 21st century in Europe. en. Journal of Geophysical Research: Biogeosciences 2013; 118. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2013JG002444:1732–47. DOI: 10.1002/2013JG002444. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/2013JG002444 [Accessed on: 2025 Oct 24]
8. Jones MW, Veraverbeke S, Andela N, Doerr SH, Kolden C, Mataveli G, Pettinari ML, Le Quéré C, Rosan TM, Werf GR van der, Wees D van, and Abatzoglou JT. Global rise in forest fire emissions linked to climate change in the extratropics. Science 2024 Oct; 386. Publisher: American Association for the Advancement of Science:eadl5889. DOI: 10.1126/science.adl5889. Available from: https://www.science.org/doi/10.1126/science.adl5889 [Accessed on: 2025 Oct 24]
9. Rovithakis A and Voulgarakis A. Wildfire aerosols and their impact on weather: A case study of the August 2021 fires in Greece using the WRF-Chem model. en. Atmospheric Science Letters 2024; 25. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/asl.1267:e1267. DOI: 10.1002/asl.1267. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/asl.1267 [Accessed on: 2025 Oct 24]
10. Driscoll DA et al. Biodiversity impacts of the 2019–2020 Australian megafires. en. Nature 2024 Nov; 635. Publisher: Nature Publishing Group:898–905. DOI: 10.1038/s41586-024-08174-6. Available from: https://www.nature.com/articles/s41586-024-08174-6 [Accessed on: 2025 Oct 24]

11. Jones BA. Are we underestimating the economic costs of wildfire smoke? An investigation using the life satisfaction approach. Journal of Forest Economics 2017; 27:80–90. DOI: 10.1016/j.jfe.2017.03.004. Available from: https://www.sciencedirect.com/science/article/pii/S1104689917300028 [Accessed on: 2025 Jul 31]
12. Bladon KD, Emelko MB, Silins U, and Stone M. Wildfire and the Future of Water Supply. Environmental Science & Technology 2014; 48:8936–43. DOI: 10.1021/es500130g. Available from: https://doi.org/10.1021/es500130g
13. Rovithakis A, Grillakis MG, Seiradakis KD, Giannakopoulos C, Karali A, Field R, Lazaridis M, and Voulgarakis A. Future climate change impact on wildfire danger over the Mediterranean: the case of Greece. en. Environmental Research Letters 2022 Apr; 17. Publisher: IOP Publishing:045022. DOI: 10.1088/1748-9326/ac5f94. Available from: https://doi.org/10.1088/1748-9326/ac5f94 [Accessed on: 2025 Oct 24]
14. Lee JH. Prediction of Large-Scale Wildfires With the Canopy Stress Index Derived from Soil Moisture Active Passive. en. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2021 Jan; 14. Publisher: IEEE:2096–102. DOI: 10.1109/JSTARS.2020.3048067. Available from: https://ieeexplore.ieee.org/document/9310300/ [Accessed on: 2025 Jul 21]
15. Shmuel A, Lazebnik T, Glickman O, Heifetz E, and Price C. Global lightning-ignited wildfires prediction and climate change projections based on explainable machine learning models. en. Scientific Reports 2025 Mar; 15. Publisher: Nature Publishing Group:7898. DOI: 10.1038/s41598-025-92171-w. Available from: https://www.nature.com/articles/s41598-025-92171-w [Accessed on: 2025 Oct 24]
16. Finney MA. The challenge of quantitative risk analysis for wildland fire. Forest Ecology and Management. Relative Risk Assessments for Decision –Making Related To Uncharacteristic Wildfire 2005 Jun; 211:97–108. DOI: 10.1016/j.foreco.2005.02.010. Available from: https://www.sciencedirect.com/science/article/pii/S0378112705000563 [Accessed on: 2025 Oct 24]
17. Sample M, Thode AE, Peterson C, Gallagher MR, Flatley W, Friggens M, Evans A, Loehman R, Hedwall S, Brandt L, Janowiak M, and Swanston C. Adaptation Strategies and Approaches for Managing Fire in a Changing Climate. en. Climate 2022 Apr; 10. Publisher: Multidisciplinary Digital Publishing Institute:58. DOI: 10.3390/cli10040058. Available from: https://www.mdpi.com/2225-1154/10/4/58 [Accessed on: 2025 Oct 24]
18. Kolström M, Lindner M, Vilén T, Maroschek M, Seidl R, Lexer MJ, Netherer S, Kremer A, Delzon S, Barbati A, Marchetti M, and Corona P. Reviewing the Science and Implementation of Climate Change Adaptation Measures in European Forestry. en. Forests 2011 Dec; 2. Publisher: Molecular Diversity Preservation International:961–82. DOI: 10.3390/f2040961. Available from: https://www.mdpi.com/1999-4907/2/4/961 [Accessed on: 2025 Oct 24]
19. Oliveras Menor I, Prat-Guitart N, Spadoni GL, Hsu A, Fernandes PM, Puig-Gironès R, Ascoli D, Bilbao BA, Bacciu V, Brotons L, Carmenta R, de-Miguel S, Gonçalves LG, Humphrey G, Ibarnegaray V, Jones MW, Machado MS, Millán A, Morais Falleiro R de, Mouillot F, Pinto C, Pons P, Regos A, Senra de Oliveira M, Harrison SP, and Armenteras Pascual D. Integrated fire management as an adaptation and mitigation strategy to altered fire regimes. en. Communications Earth & Environment 2025 Mar; 6. Publisher: Nature Publishing Group:202. DOI: 10.1038/s43247-025-02165-9. Available from: https://www.nature.com/articles/s43247-025-02165-9 [Accessed on: 2025 Oct 24]
20. Jain P, Coogan SC, Subramanian SG, Crowley M, Taylor S, and Flannigan MD. A review of machine learning applications in wildfire science and management. en. Environmental Reviews 2020 Dec; 28:478–505. DOI: 10.1139/er-2020-0019. Available from: https://cdnsciencepub.com/doi/10.1139/er-2020-0019 [Accessed on: 2025 Apr 18]
21. Di Giuseppe F, McNorton J, Lombardi A, and Wetterhall F. Global data-driven prediction of fire activity. en. Nature Communications 2025 Apr; 16. Publisher: Nature Publishing Group:2918. DOI: 10.1038/s41467-025-58097-7. Available from: https://www.nature.com/articles/s41467-025-58097-7 [Accessed on: 2025 Jul 21]

22. Xu Z, Li J, Cheng S, Rui X, Zhao Y, He H, Guan H, Sharma A, Erxleben M, Chang R, and Xu LL. Deep learning for wildfire risk prediction: Integrating remote sensing and environmental data. ISPRS Journal of Photogrammetry and Remote Sensing 2025 Sep; 227:632–77. DOI: 10.1016/j.isprsjprs.2025.06.002. Available from: https://www.sciencedirect.com/science/article/pii/S0924271625002217 [Accessed on: 2025 Oct 24]
23. Li F, Zhu Q, Yuan K, Ji F, Paul A, Lee P, Radeloff VC, and Chen M. Projecting large fires in the western US with a more trustworthy machine learning method. 2024. DOI: 10.22541/essoar.171623766.68002899/v1. Available from: https://essopenarchive.org/users/784670/articles/948622-projecting-large-fires-in-the-western-us-with-a-more-trustworthy-machine-learning-method?commit=eeeede5be0d97a799b937bd99275926bd9e8da43 [Accessed on: 2025 Apr 18]
24. Kuhn-Régnier A, Voulgarakis A, Nowack P, Forkel M, Prentice IC, and Harrison SP. The importance of antecedent vegetation and drought conditions as global drivers of burnt area. English. Biogeosciences 2021 Jun; 18. Publisher: Copernicus GmbH:3861–79. DOI: 10.5194/bg-18-3861-2021. Available from: https://bg.copernicus.org/articles/18/3861/2021/ [Accessed on: 2025 Oct 24]
25. Karasante I, Alonso L, Prapas I, Ahuja A, Carvalhais N, and Papoutsis I. SeasFire cube - a multivariate dataset for global wildfire modeling. en. Scientific Data 2025; 12:368. DOI: 10.1038/s41597-025-04546-3. Available from: https://www.nature.com/articles/s41597-025-04546-3 [Accessed on: 2025 Aug 29]
26. Prapas I, Bountos NI, Kondylatos S, Michail D, Camps-Valls G, and Papoutsis I. TeleViT: Teleconnection-driven Transformers Improve Subseasonal to Seasonal Wildfire Forecasting. 2023. DOI: 10.48550/arXiv.2306.10940. Available from: http://arxiv.org/abs/2306.10940 [Accessed on: 2025 Apr 18]
27. Kondylatos S, Prapas I, Camps-Valls G, and Papoutsis I. Mesogeos: A multi-purpose dataset for data-driven wildfire modeling in the Mediterranean. 2023. DOI: 10.48550/arXiv.2306.05144. Available from: http://arxiv.org/abs/2306.05144%20https://orionlab.space.noa.gr/mesogeos/ [Accessed on: 2025 Apr 18]
28. Kondylatos S, Prapas I, Ronco M, Papoutsis I, Camps-Valls G, Piles M, Fernández-Torres MÁ, and Carvalhais N. Wildfire Danger Prediction and Understanding With Deep Learning. en. Geophysical Research Letters 2022; 49. DOI: 10.1029/2022GL099368. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1029/2022GL099368 [Accessed on: 2025 Apr 28]
29. Abdollahi A and Pradhan B. Explainable artificial intelligence (XAI) for interpreting the contributing factors feed into the wildfire susceptibility prediction model. Science of The Total Environment 2023 Jun; 879:163004. DOI: 10.1016/j.scitotenv.2023.163004. Available from: https://www.sciencedirect.com/science/article/pii/S0048969723016224 [Accessed on: 2025 Jun 12]
30. Cilli R, Elia M, D'Este M, Giannico V, Amoroso N, Lombardi A, Pantaleo E, Monaco A, Sanesi G, Tangaro S, Bellotti R, and Lafortezza R. Explainable artificial intelligence (XAI) detects wildfire occurrence in the Mediterranean countries of Southern Europe. en. Scientific Reports 2022 Sep; 12. Publisher: Nature Publishing Group:16349. DOI: 10.1038/s41598-022-20347-9. Available from: https://www.nature.com/articles/s41598-022-20347-9 [Accessed on: 2025 Apr 29]
31. Iban MC and Sekertekin A. Machine learning based wildfire susceptibility mapping using remotely sensed fire data and GIS: A case study of Adana and Mersin provinces, Turkey. Ecological Informatics 2022 Jul; 69:101647. DOI: 10.1016/j.ecoinf.2022.101647. Available from: https://www.sciencedirect.com/science/article/pii/S1574954122000966 [Accessed on: 2025 Jun 15]
32. Bountzouklis C, Fox DM, and Bernardino ED. Predicting wildfire ignition causes in Southern France using eXplainable Artificial Intelligence (XAI) methods. en. Environmental Research Letters 2023; 18:044038. DOI: 10.1088/1748-9326/acc8ee. [Accessed on: 2025 Aug 29]

33. Zhang G, Wang M, and Liu K. Forest Fire Susceptibility Modeling Using a Convolutional Neural Network for Yunnan Province of China. en. International Journal of Disaster Risk Science 2019 Sep; 10:386–403. DOI: 10.1007/s13753-019-00233-1. Available from: https://doi.org/10.1007/s13753-019-00233-1 [Accessed on: 2025 May 7]
34. Hochreiter S and Schmidhuber J. Long Short-Term Memory. Neural Computation 1997 Nov; 9:1735–80. DOI: 10.1162/neco.1997.9.8.1735. Available from: https://ieeexplore.ieee.org/abstract/document/6795963 [Accessed on: 2025 Apr 29]
35. Lim B, Arık SÖ, Loeff N, and Pfister T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting 2021 Oct; 37:1748–64. DOI: 10.1016/j.ijforecast.2021.03.012. Available from: https://www.sciencedirect.com/science/article/pii/S0169207021000637 [Accessed on: 2025 May 12]
36. Breiman L. Random Forests. en. Machine Learning 2001 Oct; 45:5–32. DOI: 10.1023/A:1010933404324. Available from: https://doi.org/10.1023/A:1010933404324 [Accessed on: 2025 Apr 30]
37. Muñoz-Sabater J, Dutra E, Agustí-Panareda A, Albergel C, Arduini G, Balsamo G, Boussetta S, Choulga M, Harrigan S, Hersbach H, Martens B, Miralles DG, Piles M, Rodríguez-Fernández NJ, Zsoter E, Buontempo C, and Thépaut JN. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. en. Earth System Science Data 2021 Sep; 13:4349–83. DOI: 10.5194/essd-13-4349-2021. Available from: https://essd.copernicus.org/articles/13/4349/2021/ [Accessed on: 2025 Apr 18]
38. Wan Z, Hook S, and Hulley G. MOD11A1 MODIS/Terra Land Surface Temperature and the Emissivity Daily L3 Global 1km SIN Grid. NASA LP DAAC 2015
39. Didan K. MOD13A2 MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V006. 2015. DOI: 10.5067/MODIS/MOD13A2.006. Available from: https://lpdaac.usgs.gov/products/mod13a2v006/ [Accessed on: 2025 Apr 28]
40. Myneni RB, Shabanov NV, Knyazikhin Y, Yang W, Dong H, and Tan B. MOD15A2: Global LAI and FPAR. Vol. 2002. 2002 Dec :B61B–0719. Available from: https://ui.adsabs.harvard.edu/abs/2002AGUFM.B61B0719M [Accessed on: 2025 Apr 28]
41. Cammalleri C, Vogt JV, Bisselink B, and Roo A de. Comparing soil moisture anomalies from multiple independent sources over different regions across the globe. English. Hydrology and Earth System Sciences 2017 Dec; 21. Publisher: Copernicus GmbH:6329–43. DOI: 10.5194/hess-21-6329-2017. Available from: https://hess.copernicus.org/articles/21/6329/2017/hess-21-6329-2017.html [Accessed on: 2025 Apr 28]
42. Tatem AJ. WorldPop, open data for spatial demography. en. Scientific Data 2017 Jan; 4. Publisher: Nature Publishing Group:170004. DOI: 10.1038/sdata.2017.4. Available from: https://www.nature.com/articles/sdata20174 [Accessed on: 2025 Apr 28]
43. Franks S and Rengarajan R. Evaluation of Copernicus DEM and Comparison to the DEM Used for Landsat Collection-2 Processing. en. Remote Sensing 2023 Jan; 15. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute:2509. DOI: 10.3390/rs15102509. Available from: https://www.mdpi.com/2072-4292/15/10/2509 [Accessed on: 2025 Apr 28]
44. Potapov P, Hansen MC, Pickens A, Hernandez-Serna A, Tyukavina A, Turubanova S, Zalles V, Li X, Khan A, Stolle F, Harris N, Song XP, Baggett A, Kommareddy I, and Kommareddy A. The Global 2000-2020 Land Cover and Land Use Change Dataset Derived From the Landsat Archive: First Results. English. Frontiers in Remote Sensing 2022 Apr; 3. Publisher: Frontiers. DOI: 10.3389/frsen.2022.856903. Available from: https://www.frontiersin.orghttps://www.frontiersin.org/journals/remote-sensing/articles/10.3389/frsen.2022.856903/full [Accessed on: 2025 Apr 28]
45. Giglio L, Schroeder W, and Justice CO. The collection 6 MODIS active fire detection algorithm and fire products. Remote Sensing of Environment 2016 Jun; 178:31–41. DOI: 10.1016/j.rse.2016.02.054. Available from: https://www.sciencedirect.com/science/article/pii/S0034425716300827 [Accessed on: 2025 Apr 28]

46. Huot F, Hu RL, Ihme M, Wang Q, Burge J, Lu T, Hickey J, Chen YF, and Anderson J. Deep Learning Models for Predicting Wildfires from Historical Remote-Sensing Data. 2021. DOI: 10.48550/arXiv.2010.07445. Available from: http://arxiv.org/abs/2010.07445 [Accessed on: 2025 May 7]
47. BergstraJames and BengioYoshua. Random search for hyper-parameter optimization. EN. The Journal of Machine Learning Research 2012 Feb. DOI: 10.5555/2188385.2188395. Available from: https://dl.acm.org/doi/10.5555/2188385.2188395 [Accessed on: 2025 Oct 1]
48. Liashchynskyi P and Liashchynskyi P. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. en. 2019 Dec. Available from: https://arxiv.org/abs/1912.06059v1 [Accessed on: 2025 Oct 2]
49. Li F, Zhu Q, Riley WJ, Zhao L, Xu L, Yuan K, Chen M, Wu H, Gui Z, Gong J, and Randerson JT. AttentionFire_v1.0: interpretable machine learning fire model for burned-area predictions over tropics. English. Geoscientific Model Development 2023 Feb; 16. Publisher: Copernicus GmbH:869–84. DOI: 10.5194/gmd-16-869-2023. Available from: https://gmd.copernicus.org/articles/16/869/2023/ [Accessed on: 2025 Oct 2]
50. Akiba T, Sano S, Yanase T, Ohta T, and Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. New York, NY, USA: Association for Computing Machinery, 2019 :2623–31. DOI: 10.1145/3292500.3330701. Available from: https://dl.acm.org/doi/10.1145/3292500.3330701 [Accessed on: 2025 May 14]
51. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 1958; 65:386–408. DOI: 10.1037/h0042519
52. Cho K, Merrienboer Bv, Bahdanau D, and Bengio Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. 2014. DOI: 10.48550/arXiv.1409.1259. Available from: http://arxiv.org/abs/1409.1259 [Accessed on: 2025 Apr 30]
53. Krizhevsky A, Sutskever I, and Hinton GE. ImageNet classification with deep convolutional neural networks. Commun. ACM 2017 May; 60:84–90. DOI: 10.1145/3065386. Available from: https://dl.acm.org/doi/10.1145/3065386 [Accessed on: 2025 Jul 26]
54. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Łu, and Polosukhin I. Attention is All you Need. *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html [Accessed on: 2025 May 9]
55. Harris L and Grzes M. Comparing Explanations between Random Forests and Artificial Neural Networks. *IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 2019 :2978–85. DOI: 10.1109/SMC.2019.8914321. [Accessed on: 2025 Aug 1]
56. Lek S and Park YS. Multilayer Perceptron. *Encyclopedia of Ecology*. Ed. by Jørgensen SE and Fath BD. Oxford: Academic Press, 2008 Jan :2455–62. DOI: 10.1016/B978-008045405-4.00162-2. Available from: https://www.sciencedirect.com/science/article/pii/B9780080454054001622 [Accessed on: 2025 Oct 8]
57. Sasaki Y. The truth of the F-measure. en. 2007; 1:5. Available from: https://nicolasshu.com/assets/pdf/Sasaki_2007_The%20Truth%20of%20the%20F-measure.pdf
58. Raschka S. An Overview of General Performance Metrics of Binary Classifier Systems. 2014 Oct. DOI: 10.48550/arXiv.1410.5330. Available from: http://arxiv.org/abs/1410.5330 [Accessed on: 2025 Oct 1]
59. Bommer PL, Kretschmer M, Hedström A, Bareeva D, and Höhne MMC. Finding the Right XAI Method—A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science. EN. Artificial Intelligence for the Earth Systems 2024 Jun; 3. Publisher: American Meteorological Society Section: Artificial Intelligence for the Earth Systems. DOI: 10.1175/AIES-D-23-0074.1. Available from: https://journals.ametsoc.org/view/journals/aies/3/3/AIES-D-23-0074.1.xml [Accessed on: 2025 Nov 3]

60. Lundberg SM and Lee SI. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. Available from: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html [Accessed on: 2025 Apr 25]

61. Molnar C. Interpretable Machine Learning - A Guide for Making Black Box Models Explainable. 3rd ed. 2025. Available from: https://christophm.github.io/interpretable-ml-book

62. Abdollahi A and Pradhan B. Urban Vegetation Mapping from Aerial Imagery Using Explainable AI (XAI). en. Sensors 2021 Jan; 21. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute:4738. DOI: 10.3390/s21144738. Available from: https://www.mdpi.com/1424-8220/21/14/4738 [Accessed on: 2025 Jun 12]

63. Sundararajan M, Taly A, and Yan Q. Axiomatic Attribution for Deep Networks. en. *International Conference on Machine Learning*. PMLR, 2017 :3319–28. [Accessed on: 2025 Apr 26]

64. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, and Desmaison A. Automatic differentiation in PyTorch. en. 2017

65. Weng X, Forster GL, and Nowack P. A machine learning approach to quantify meteorological drivers of ozone pollution in China from 2015 to 2019. English. Atmospheric Chemistry and Physics 2022 Jun; 22. Publisher: Copernicus GmbH:8385–402. DOI: 10.5194/acp-22-8385-2022. Available from: https://acp.copernicus.org/articles/22/8385/2022/ [Accessed on: 2025 Oct 1]

66. Hickman SHM, Griffiths PT, Nowack PJ, and Archibald AT. Short-term forecasting of ozone air pollution across Europe with transformers. en. Environmental Data Science 2023 Jan; 2:e43. DOI: 10.1017/eds.2023.37. Available from: https://www.cambridge.org/core/journals/environmental-data-science/article/shortterm-forecasting-of-ozone-air-pollution-across-europe-with-transformers/9A30B0A8F17FC59DF55FAA0401474AED [Accessed on: 2025 Oct 1]

67. Kruskal WH and Wallis WA. Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association 1952 Dec; 47:583–621. DOI: 10.1080/01621459.1952.10483441. Available from: https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441 [Accessed on: 2025 Jul 3]

68. Ramachandra V. Artificial Intelligence in Climate Science: A State-of-the-Art Review (2020–2025). en. 2025 Jul. Publisher: EarthArXiv. Available from: https://eartharxiv.org/repository/view/9661/ [Accessed on: 2025 Oct 9]

69. Castrejon DJ, Wang C, Osmak D, Kukadiya B, Liu L, Giraldo M, and Jiang X. Machine Learning-based California Wildfire Risk Prediction and Visualization. *2023 International Conference on Machine Learning and Applications (ICMLA)*. ISSN: 1946-0759. 2023 Dec :1212–7. DOI: 10.1109/ICMLA58977.2023.00182. Available from: https://ieeexplore.ieee.org/document/10459942/ [Accessed on: 2025 Oct 9]

70. Schmitt EA, Zaremba E, Ananthavaram N, Liu L, Giraldo M, and Jiang X. Ecosystem-Based Wildfire Risk Prediction with Machine Learning. *2024 IEEE International Conference on Big Data (BigData)*. ISSN: 2573-2978. 2024 Dec :7540–5. DOI: 10.1109/BigData62323.2024.10825794. Available from: https://ieeexplore.ieee.org/document/10825794/ [Accessed on: 2025 Oct 9]

71. Rosner B, Glynn RJ, and Lee MLT. The Wilcoxon Signed Rank Test for Paired Comparisons of Clustered Data. Biometrics 2006 Mar; 62:185–92. DOI: 10.1111/j.1541-0420.2005.00389.x. Available from: https://doi.org/10.1111/j.1541-0420.2005.00389.x [Accessed on: 2025 Sep 24]

72. Pölz A, Blaschke AP, Komma J, Farnleitner AH, and Derx J. Transformer Versus LSTM: A Comparison of Deep Learning Models for Karst Spring Discharge Forecasting. en. Water Resources Research 2024; 60. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022WR032602:e2022WR032602. DOI: 10.1029/2022WR032602. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1029/2022WR032602 [Accessed on: 2025 Oct 2]

73. Michail D, Panagiotou LI, Davalas C, Prapas I, Kondylatos S, Bountos NI, and Papoutsis I. Seasonal Fire Prediction using Spatio-Temporal Deep Neural Networks. 2024. DOI: 10.48550/arXiv.2404.06437. Available from: http://arxiv.org/abs/2404.06437 [Accessed on: 2025 Apr 18]

74. Sturmfels P, Lundberg S, and Lee SI. Visualizing the Impact of Feature Attribution Baselines. en. Distill 2020 Jan; 5:e22. DOI: 10.23915/distill.00022. Available from: https://distill.pub/2020/attribution-baselines [Accessed on: 2025 Nov 4]

75. Debeire K, Bock L, Nowack P, Runge J, and Eyring V. Constraining uncertainty in projected precipitation over land with causal discovery. English. Earth System Dynamics 2025 Apr; 16. Publisher: Copernicus GmbH:607–30. DOI: 10.5194/esd-16-607-2025. Available from: https://esd.copernicus.org/articles/16/607/2025/ [Accessed on: 2025 Oct 24]

76. Nowack P, Runge J, Eyring V, and Haigh JD. Causal networks for climate model evaluation and constrained projections. en. Nature Communications 2020 Mar; 11. Publisher: Nature Publishing Group:1415. DOI: 10.1038/s41467-020-15195-y. Available from: https://www.nature.com/articles/s41467-020-15195-y [Accessed on: 2025 Oct 24]

77. Hickman S, Trajkovic I, Kaltenborn J, Pelletier F, Archibald A, Gurwicz Y, Nowack P, Rolnick D, and Boussard J. Causal Climate Emulation with Bayesian Filtering. arXiv:2506.09891 [cs]. 2025 Jun. DOI: 10.48550/arXiv.2506.09891. Available from: http://arxiv.org/abs/2506.09891 [Accessed on: 2025 Oct 24]

78. Aas K, Jullum M, and Løland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence 2021 Sep; 298:103502. DOI: 10.1016/j.artint.2021.103502. Available from: https://www.sciencedirect.com/science/article/pii/S0004370221000539 [Accessed on: 2025 Oct 17]

79. Wilkinson S, Nowack P, and Joshi M. Process-Based Machine Learning Observationally Constrains Future Regional Warming Projections. en. Journal of Geophysical Research: Machine Learning and Computation 2025; 2. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2025JH000698:e2025JH000698. DOI: 10.1029/2025JH000698. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1029/2025JH000698 [Accessed on: 2025 Oct 24]

80. Huang X and Marques-Silva J. On the failings of Shapley values for explainability. International Journal of Approximate Reasoning. Synergies between Machine Learning and Reasoning 2024 Aug; 171:109112. DOI: 10.1016/j.ijar.2023.109112. Available from: https://www.sciencedirect.com/science/article/pii/S0888613X23002438 [Accessed on: 2025 Oct 5]

81. Chuvieco E, Yebra M, Martino S, Thonicke K, Gómez-Giménez M, San-Miguel J, Oom D, Velea R, Mouillot F, Molina JR, Miranda AI, Lopes D, Salis M, Bugaric M, Sofiev M, Kadantsev E, Gitas IZ, Stavrakoudis D, Eftychidis G, Bar-Massada A, Neidermeier A, Pampanoni V, Pettinari ML, Arrogante-Funes F, Ochoa C, Moreira B, and Viegas D. Towards an Integrated Approach to Wildfire Risk Assessment: When, Where, What and How May the Landscapes Burn. en. Fire 2023 May; 6:215. DOI: 10.3390/fire6050215. Available from: https://www.mdpi.com/2571-6255/6/5/215 [Accessed on: 2025 Jun 11]

82. Bistinas I, Oom D, Sá ACL, Harrison SP, Prentice IC, and Pereira JMC. Relationships between Human Population Density and Burned Area at Continental and Global Scales. en. PLOS ONE 2013 Dec; 8. Publisher: Public Library of Science:e81188. DOI: 10.1371/journal.pone.0081188. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0081188 [Accessed on: 2025 Aug 5]

83. Li LM, Song WG, Ma J, and Satoh K. Artificial neural network approach for modeling the impact of population density and weather parameters on forest fire risk. en. International Journal of Wildland Fire 2009 Sep; 18. Publisher: CSIRO PUBLISHING:640–7. DOI: 10.1071/WF07136. Available from: https://www.publish.csiro.au/wf/WF07136 [Accessed on: 2025 Aug 5]

84. Copernicus Emergency Management Service data. Devastating wildfire in Sierra de la Culebra, Spain — Copernicus. 2022. Available from: https://www.copernicus.eu/en/media/image-day-gallery/devastating-wildfire-sierra-de-la-culebra-spain [Accessed on: 2025 Jul 1]

85. Alonso L, Gans F, Karasante I, Ahuja A, Prapas I, Kondylatos S, Papoutsis I, Panagiotou E, Mihail D, Cremer F, Weber U, and Carvalhais N. SeasFire Cube: A Global Dataset for Seasonal Fire Modeling in the Earth System. 2024. DOI: 10.5281/ZENODO.13834057. Available from: https://zenodo.org/doi/10.5281/zenodo.13834057 [Accessed on: 2025 Apr 18]

86. Perr-Sauer J, Ugirumurera J, Gafur J, Bensen EA, Nguyen T, Paul S, Severino J, Nag A, Vijayshankar S, Gasper P, Finegan DP, Holden J, Mueller J, Graf P, Tripp C, and Egan H. Applications of explainable artificial intelligence in renewable energy research. Energy Reports 2025 Dec; 14:2217–35. DOI: 10.1016/j.egyr.2025.08.046. Available from: https://www.sciencedirect.com/science/article/pii/S2352484725005116 [Accessed on: 2025 Oct 24]

87. Vasconcelos RN, Santana MMM de, Costa DP, Duverger SG, Ferreira-Ferreira J, Oliveira M, Barbosa LdS, Cordeiro CL, and Franca Rocha WJS. Machine Learning Model Reveals Land Use and Climate's Role in Caatinga Wildfires: Present and Future Scenarios. en. Fire 2025 Jan; 8. Publisher: Multidisciplinary Digital Publishing Institute:8. DOI: 10.3390/fire8010008. Available from: https://www.mdpi.com/2571-6255/8/1/8 [Accessed on: 2025 Oct 24]

88. Bhattarai H, Val Martin M, Sitch S, Yung DHY, and Tai APK. Global patterns and drivers of climate-driven fires in a warming world. en. 2025 Mar. DOI: 10.5194/egusphere-2025-804. Available from: https://egusphere.copernicus.org/preprints/2025/egusphere-2025-804/ [Accessed on: 2025 Oct 24]

89. Roscher R, Bohn B, Duarte MF, and Garcke J. Explainable Machine Learning for Scientific Insights and Discoveries. IEEE Access 2020; 8:42200–16. DOI: 10.1109/ACCESS.2020.2976199. Available from: https://ieeexplore.ieee.org/document/9007737 [Accessed on: 2025 Oct 17]

90. Mukunga T, Forkel M, Forrest M, Zotta RM, Pande N, Schlaffer S, and Dorigo W. Effect of Socioeconomic Variables in Predicting Global Fire Ignition Occurrence. en. Fire 2023 May; 6. Publisher: Multidisciplinary Digital Publishing Institute:197. DOI: 10.3390/fire6050197. Available from: https://www.mdpi.com/2571-6255/6/5/197 [Accessed on: 2025 Oct 9]

91. DeFries RS, Morton DC, Werf GR van der, Giglio L, Collatz GJ, Randerson JT, Houghton RA, Kasibhatla PK, and Shimabukuro Y. Fire-related carbon emissions from land use transitions in southern Amazonia. en. Geophysical Research Letters 2008; 35. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2008GL035689. DOI: 10.1029/2008GL035689. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1029/2008GL035689 [Accessed on: 2025 Oct 9]

92. Bistinas I, Harrison SP, Prentice IC, and Pereira JMC. Causal relationships versus emergent patterns in the global controls of fire frequency. English. Biogeosciences 2014 Sep; 11. Publisher: Copernicus GmbH:5087–101. DOI: 10.5194/bg-11-5087-2014. Available from: https://bg.copernicus.org/articles/11/5087/2014/ [Accessed on: 2025 Oct 9]

93. Knorr W, Kaminski T, Arneth A, and Weber U. Impact of human population density on fire frequency at the global scale. English. Biogeosciences 2014 Feb; 11. Publisher: Copernicus GmbH:1085–102. DOI: 10.5194/bg-11-1085-2014. Available from: https://bg.copernicus.org/articles/11/1085/2014/bg-11-1085-2014.html?utm_source=chatgpt.com [Accessed on: 2025 Oct 9]

94. Andela N and Werf GR van der. Recent trends in African fires driven by cropland expansion and El Niño to La Niña transition. en. Nature Climate Change 2014 Sep; 4. Publisher: Nature Publishing Group:791–5. DOI: 10.1038/nclimate2313. Available from: https://www.nature.com/articles/nclimate2313 [Accessed on: 2025 Oct 9]

95. Zhao H, Zhang Z, Wang X, Zhen S, Zhang H, Bu ZJ, Zhao J, Guo X, Wei K, and Dong L. Future enhanced threshold effects of wildfire drivers could increase burned areas in northern mid- and high latitudes. en. Communications Earth & Environment 2025 Mar; 6. Publisher: Nature Publishing Group:224. DOI: 10.1038/s43247-025-02202-7. Available from: https://www.nature.com/articles/s43247-025-02202-7 [Accessed on: 2025 Jun 16]

96. Xu Z, Li J, Cheng S, Rui X, Zhao Y, He H, and Xu L. Wildfire Risk Prediction: A Review. 2024 Oct. DOI: 10.48550/arXiv.2405.01607. Available from: http://arxiv.org/abs/2405.01607 [Accessed on: 2025 Jun 23]

97. Radeloff VC, Mockrin MH, Helmers D, Carlson A, Hawbaker TJ, Martinuzzi S, Schug F, Alexandre PM, Kramer HA, and Pidgeon AM. Rising wildfire risk to houses in the United States, especially in grasslands and shrublands. Science 2023 Nov; 382. Publisher: American Association for the Advancement of Science:702–7. DOI: 10.1126/science.ade9223. Available from: https://www.science.org/doi/10.1126/science.ade9223 [Accessed on: 2025 Jun 23]

98. Holsten A, Dominic AR, Costa L, and Kropp JP. Evaluation of the performance of meteorological forest fire indices for German federal states. en. Forest Ecology and Management 2013 Jan; 287:123–31. DOI: 10.1016/j.foreco.2012.08.035. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0378112712005191 [Accessed on: 2025 Jun 11]

99. Pimont F, Dupuy JL, and Linn RR. Coupled slope and wind effects on fire spread with influences of fire size: a numerical study using FIRETEC. en. International Journal of Wildland Fire 2012 Jul; 21. Publisher: CSIRO PUBLISHING:828–42. DOI: 10.1071/WF11122. Available from: https://www.publish.csiro.au/wf/WF11122 [Accessed on: 2025 Jun 23]

100. Weise DR and Biging GS. A Qualitative Comparison of Fire Spread Models Incorporating Wind and Slope Effects. Forest Science 1997 May; 43:170–80. DOI: 10.1093/forestscience/43.2.170. Available from: https://doi.org/10.1093/forestscience/43.2.170 [Accessed on: 2025 Jun 23]

101. Copernicus Sentinel-2 imagery. Spanish Province of Zamora, in Castilla y León, ravaged by wildfires — Copernicus. Article. 2022. Available from: https://www.copernicus.eu/en/media/image-day-gallery/spanish-province-zamora-castilla-y-leon-ravaged-wildfires [Accessed on: 2025 Jul 1]