PAPER • **OPEN ACCESS**

Assessing and explaining temporal deep learning models for wildfire danger prediction

To cite this article: Pauline Becker *et al* 2026 *Mach. Learn.: Earth* **2** 015014

MACHINE LEARNING
Earth

OPEN ACCESS

RECEIVED
6 November 2025

REVISED
13 March 2026

ACCEPTED FOR PUBLICATION
2 April 2026

PUBLISHED
28 April 2026

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
[*] Author to whom any correspondence should be addressed.

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

Keywords: wildfires, Mediterranean, explainable AI, SHAP, Transformers, AI in climate science

Supplementary material for this article is available online

Abstract

Modern methods for wildfire danger prediction are critical for mitigating the detrimental impacts of fires on ecosystems, public health, and the economy. While Machine Learning has emerged as a powerful approach to model the complex interactions driving wildfire risk, its 'black-box' nature creates a trade-off between predictive skill and physical plausibility and interpretability required for trustworthy risk assessments. In this study, we systematically assess the predictive performance and physical consistency of seven temporal deep learning (DL) models against two decision tree-based baselines, random forest (RF) and XGBoost (XGB), for next-day wildfire danger prediction in the Mediterranean. We apply explainable AI (xAI) methods to interpret model attributions and assess their broad alignment with established fire science. Results show that all DL models outperform RF and XGB baselines, with Transformer models achieving the highest predictive accuracy (F_1-score > 0.81), significantly outperforming the RF baseline (post-hoc Dunn test, $p < 10^{-5}$) and by effectively capturing long-range temporal dependencies. However, xAI analyses reveal a key trade-off: despite their higher predictive performance, DL models exhibit lower physical consistency in their averaged driver relationships. Specifically, when evaluated against 19 expected fire-driver relationships, the RF and XGB correctly capture 13 (12) relationships, whereas DL models capture at most 11. 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 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 [2–5]. Beyond their implications for carbon emissions [4, 6–8], 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 [14–16], 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 [20–22]. Recent studies show that ML algorithms can outperform conventional fire weather indices globally [21] and in regional (e.g. over the Western US) high-resolution wildfire prediction tasks [23]. In particular, gradient boosting approaches such as XGBoost (XGB) have demonstrated strong and consistent predictive skill across diverse climatic and geographic settings [24–29]. 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 [30], 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 [31], for forecasting global burned area patterns across multiple temporal windows at coarse spatial resolutions. Kondylatos *et al* [32], in turn, introduced the Mesogeos fire cube [32], 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* [24] 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 model explanations is increasingly recognized as a central challenge in the application of ML 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.

xAI is becoming integral to wildfire modeling, as a means of understanding ML predictions. Post-hoc techniques, particularly SHAP, are now routinely applied to elucidate driver importance and model behavior [25, 33–35]. However, this growing adoption remains largely confined to single model-xAI pairings [33, 34, 36] or focus on tree-based or shallow ML architectures [25–29]. While initial efforts to assess and interpret DL models have begun to emerge (e.g. [24, 37]), they remain comparatively underexplored and insufficiently benchmarked. This gap is particularly pressing as temporal DL architectures offer methodological advantages for next-day wildfire danger prediction in the emerging 'big-data/big-model' era for observational wildfire datasets. These models are inherently suited to automatically learn complex temporal dependencies in Earth system problems [38]. Here, we conduct a systematic, benchmarked inter-comparison of multiple temporal DL architectures. Using a fixed, large-scale dataset and consistent evaluation criteria, we assess not only overall predictive performance but also the physical consistency of model reasoning through xAI metrics. We hope this dual evaluation framework will provide a solid foundation to inform future wildfire modeling with DL.

2. Data and methods

2.1. Data and preprocessing

In this study, we use the Mesogeos data cube [32] 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, dew-point temperature, surface pressure, relative humidity (rh), precipitation, and surface solar radiation, derived from the ERA5-Land dataset [39]. Vegetation status and land surface conditions were represented using daytime and nighttime land surface temperature [40], the normalized difference vegetation index (NDVI) [41], and the leaf area index (LAI) [42] from MODIS, alongside soil moisture estimates from the European drought observatory [43]. Indicators of human presence and activity, such as population

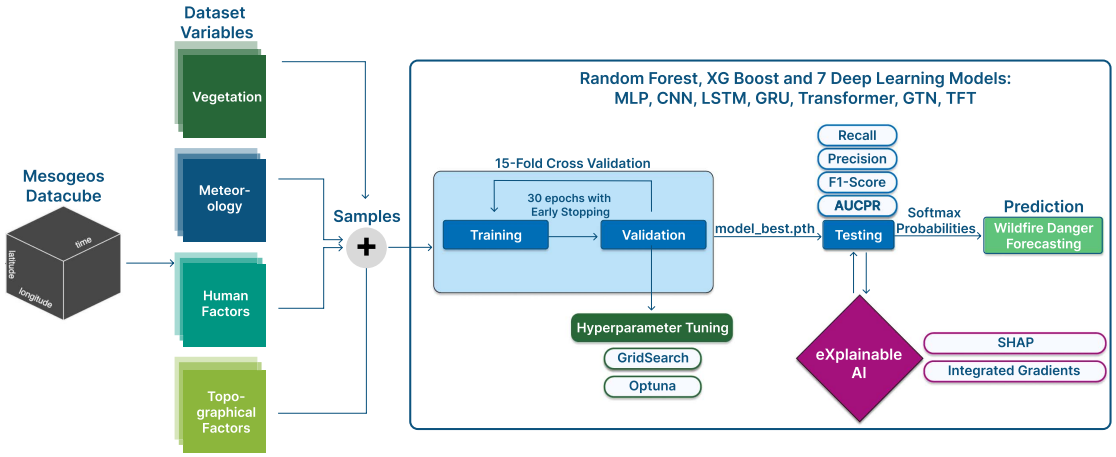

Figure 1. Overview of the ML pipeline. Samples extracted from the Mesogeos [32] data cube are used to train, validate, and test XGB, RF and DL models via a cross-validation scheme. Final models are evaluated on the test set and interpreted using xAI techniques.

density and proximity to roads, were obtained from WorldPop [44]. Terrain characteristics, including elevation and slope, were incorporated using the COP-DEM dataset [45], while land cover classifications were sourced from the Copernicus Climate Change Service [46]. Burned areas are from EFFIS [47], while ignition points and ignition dates were estimated using the MODIS Active Fire product [48]. A full list of all Mesogeos predictor variables, their abbreviations, categories, and units is provided in table S1 in the supplementary material.

We preprocessed the data following the Mesogeos definitions [32], 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 d 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 reduce the likelihood of selecting unburned locations that nonetheless exhibited high fire danger [32]. We sampled twice as many negative samples as positives ones from the overall dataset, following standard practices [24, 32, 49, 50], such that the negative samples roughly match the land-cover distribution 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 [23, 51–53] and Optuna [54], exploring parameters such as learning rate, batch size, dropout rate, and weight decay. The final configurations were selected based on the validation F_1-score. The final hyperparameter values for each model are reported in table S3.

We systematically compared seven DL architectures, including multilayer perceptron (MLP) [55], long short-term memory (LSTM) [56], gated recurrent unit (GRU) [57], convolutional neural network (CNN) [58], Transformer [59], gated Transformer network (GTN) [32], and temporal fusion Transformer (TFT) [60]. These models were evaluated against two established tree-based models, random forest (RF) [61] and XGB [62]. RF and XGB were selected as baselines due to their consistently strong performance in wildfire prediction across diverse geographic and climatic settings [24–29]. This comparison was motivated by the need to evaluate whether increasing model complexity translates into superior predictive power and/or physical consistency for wildfire danger predictions, 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 intrinsic temporal awareness [63]. XGB uses regularized gradient-boosted decision trees to capture complex non-linear

effects and interactions [62, 64]. MLPs, or feed-forward neural networks, are a standard neural network architecture to learn non-linear interactions between input and outputs, subject to one or more hidden layers with variable numbers of hidden neurons. However, MLPs have no intrinsic architectural component to efficiently exploit temporal or spatial structure [65]. CNNs [58, e.g.], in turn, are designed to mimic the biological receptive field by learning shared kernel weights, or filters, to automatically detect and abstract patterns in structured image data. CNNs have been adapted to multivariate time-series forecasting by reshaping sequences into pseudo-images (e.g. time as 'height,' features as 'width' [66, 67]), allowing shared-weight convolutions to extract temporal patterns. As an improvement over vanilla recurrent neural networks (RNNs) [68], LSTMs [56] introduce cell states and gating mechanisms to also retain long-term temporal information, while GRUs simplify this structure, achieving similar performance at lower computational cost [57]. Finally, the transformer architecture enables modeling of long-range dependencies independent of the distance in sequences (thus overcoming still remaining long-distance problems of LSTMs) while simultaneously facilitating parallelization compared to sequential computations and backpropagation in RNNs [59], Later extensions, such as the GTN and TFT, further aim to enhance specific strengths of the vanilla transformer. GTN incorporates the attention not only across input variables but also across time, which may enable the model to capture complex interactions of the variables [32], whereas TFT explicitly considers static and dynamic inputs through dedicated gating mechanisms, making it the furthest augmented architecture considered in our study [60].

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.

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 [69], which is well-suited for imbalanced classification where wildfire occurrences are rare compared to non-fire or negative events [70]. The F_1-score (equation (1)) is the harmonic mean of precision and recall, balancing false positives and false negatives:

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

In addition to the F_1-score, we also report accuracy (equation (2)), precision (equation (3)) and recall (equation (4)) to explicitly quantify the overall fraction of correctly classified samples, the reliability of positive predictions, and the ability to detect positive samples, respectively,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4}$$

Here, TP (true positives) denotes correctly predicted fire-danger events, TN (true negatives) correctly predicted non-fire-danger events, FP (false positives) non-fire-danger events incorrectly predicted as fire danger, and FN (false negatives) fire-danger events that were not detected by the model. All four metrics range from 0 to 1, with values closer to 1 indicating better predictive performance. Furthermore, we include the area under the precision-recall curve (AUPRC) as a threshold-independent metric that also ranges from 0–1 [71]. AUPRC provides an indication of how well the positive and negative classes are separated [71], with higher values reflecting stronger discriminative performance.

2.4. xAI

Despite their overall high predictive performance, 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 explainable methods in climate science [72], this study employs attribution-based xAI techniques to uncover how input features shape fire predictions.

Different xAI methods rely on distinct assumptions and approximation strategies to estimate feature influence, and therefore will yield at least somewhat different attribution patterns for the same trained model. In this study, we employ two xAI approaches, SHAP values and integrated gradients (IGs), to help mitigate the limitations of any single method. However, we advise readers to interpret these results with caution, as post-hoc xAI methods have inherent limitations that are beyond the scope of this study to address. For further discussion of these limitations, we refer the reader to [73].

To ensure a consistent cross-architecture explainability assessment, we apply SHAP uniformly across all model classes, including both DL architectures and tree-based baselines (RF and XGBoost). IG is used as a complementary method specifically for differentiable models, providing additional validation of attribution patterns within the DL models. 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 [74, 75] or the intrinsic interpretability of attention weights in Transformers [76]. However, these methods are architecture-specific and therefore not suitable for our study, which spans heterogeneous model classes.

SHAP values [77], 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 [78]. In practice, we applied Kernel SHAP, a model-agnostic approximation method [79] implemented in the SHAP Python library [77], which balances accuracy and computational cost by requiring fewer model evaluations than other sampling-based approaches [80]. To ensure stable attributions in this high-dimensional setting (30 d * 24 features = 720 inputs), Kernel SHAP was computed using a background set of 100 randomly selected test samples and n_samples = 1000 coalition samples per explained instance. Computing SHAP values for the full test set ($N = 4107$) required approximately 2.4 h wall-clock time on an NVIDIA A100 (40 GB). For tree-based baselines (RF and XGB), SHAP values were obtained via TreeExplainer contributions [81], which do not require coalition sampling and were computed for the full test set in approximately 1.1 h wall-clock time on a CPU-only node.

IGs [82] 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 [83] 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.

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 table 1 and figure S1 in the supplementary material). To formally test performance differences among models, we applied both parametric (ANOVA F-test) and non-parametric (Kruskal–Wallis [84]) analyses. For the F_1-score, both tests rejected the null hypothesis of equal model performance ($p < 10^{-14}$). 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 [23, 24] in which DL models outperformed the RF baseline. For the XGB baseline, post-hoc comparisons revealed smaller and less consistent performance differences, with only partial statistical significance across DL models (see figure S6 for detailed results).

Beyond the F_1-score, differences in Accuracy and Recall were also statistically significant under both ANOVA and Kruskal–Wallis tests ($p < 10^{-4}$). In contrast, no statistically significant differences were

Table 1. AUPRC, Precision, Recall, F_1-Score, and Accuracy test-set performance averaged over cross-validation runs for all model classes. Overall, XGB and RF show consistently lower performance across the reported metrics compared to the DL models. For each metric, the score of the best performing model is highlighted in bold.

Model	AUPRC	Precision	Recall	F_1	Accuracy
Transformer	**0.889**	0.775	**0.866**	**0.813**	**0.871**
GTN	0.873	0.762	0.856	0.809	0.862
MLP	0.878	**0.789**	0.833	0.805	0.64
LSTM	0.876	0.769	0.844	0.805	0.864
TFT	0.877	0.779	0.827	0.801	0.860
GRU	0.874	0.772	0.843	0.800	0.861
CNN	0.857	0.756	0.829	0.788	0.851
XGB	0.872	0.765	0.815	0.786	0.852
RF	0.841	0.785	0.724	0.753	0.841

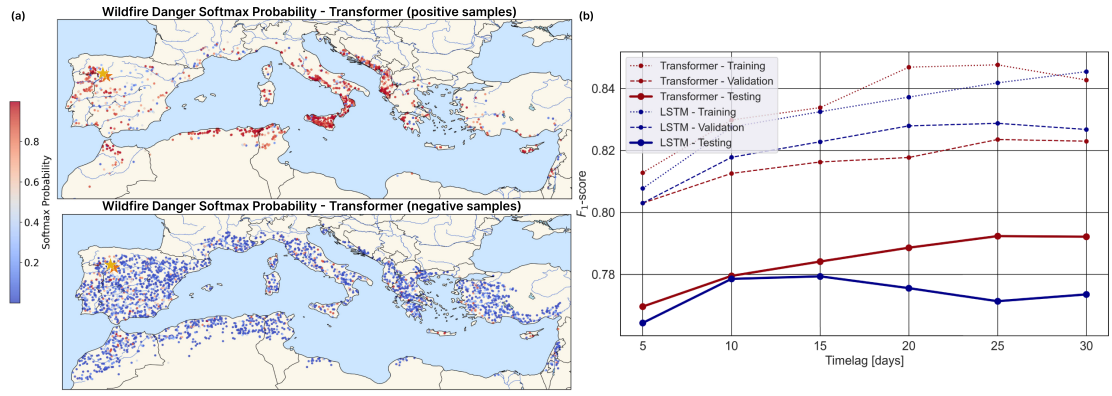

Figure 2. Spatial and temporal evaluation of model behavior. (a) Softmax probabilities predicted by the Transformer for positive (top) and negative (bottom) test samples, where red indicates high fire danger and blue indicates low fire danger. The two stars (in yellow and orange) mark the locations of 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 d) 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.

observed for Precision, with both parametric and non-parametric tests yielding $p > 0.05$, indicating that none of the evaluated architectures demonstrated a clear advantage in reducing false positive predictions.

Among the DL architectures, the Transformer achieved the best performance in four out of five reported metrics, including AUPRC, Recall, F_1-score, and Accuracy (table 1). Additionally, subsequent analyses revealed that Transformer models maintained performance particularly well under varying temporal input windows (see section 3.3). However, no statistically significant differences were detected among the DL architectures. 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 [32], our re-implementation of baseline architectures (LSTM, Transformer, GTN) achieved slightly higher F_1-scores through cross-validation and extensive hyperparameter tuning, for instance improving the Transformer model from a previously reported F_1-score of 0.78–0.81 in our experiments. Incorporating additional architectures further increased overall performance, with nearly all DL models reaching or exceeding F_1-scores values of 0.78. Consistent improvements are also observed relative to prior wildfire-danger studies focusing on Greece and surrounding regions. In [85], RF and CNN reached $F_1 = 0.631$ and $F_1 \approx 0.63$, respectively, while [24] reported RF performance of $F_1 = 0.703$; in contrast, our cross-validated results yield $F_1 = 0.753$ for RF and $F_1 = 0.788$ for CNN.

3.2. Spatial evaluation of model performance

To complement quantitative performance metrics, we visualized the spatial distribution of softmax probabilities predicted by the best-performing model (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 2, 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

Table 2. 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 F_1-scores exceed inland, though the average coastal-inland gap is smaller with two separate models.

	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
XGB	0.81	0.60	0.82	0.66

in coastal versus inland regions, may influence model performance, reflecting a common challenge in ML [86–88].

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 on the test data to quantify the coastal-inland effect. Across all architectures, coastal performance exceeded inland performance (figure S2). Among the DL models, coastal-inland differences were consistently pronounced (on average $\Delta F_1 \geqslant 0.17$), 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 nine 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 [89], show that coastal predictions achieve significantly higher F_1 scores than inland predictions (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 available for training, we tested a region-specific modeling strategy by subdividing the dataset and training separate models for coastal and inland areas (table 2). This idea is similar to the approach followed by Schmitt *et al* [88], 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 $\Delta F_1 \approx 0.04$ across models relative to the one-model baseline, with the Transformer showing the largest gain ($\Delta F_1 = 0.11$). 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 for the case of two separate models, we can still see the systematic effects in reduced inland performance as the coastal-inland gap remains substantial in the two-model case (mean $\Delta F_1 \approx 0.07$), even under region-specific training. An exception was the RF, whose inland F_1-score decreased by $\Delta F_1 = -0.23$ when trained solely on inland samples, suggesting it benefits primarily from larger overall sample sizes regardless of spatial origin.

3.3. Performance dependence on temporal context length

We further evaluated the impact of historical input length (5–30 d) on predictive performance (figure 2(b)). Across all temporal input lengths, the Transformer consistently outperformed the LSTM. Its performance increased nearly linearly with longer input sequences up to about 25 d, after which it began to plateau. This result demonstrates the advantage of attention-based models in capturing longer

range temporal dependencies [59, 90]. In contrast, LSTM performance saturated after around 15 d 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 [76]. These results support prior studies [31, 91], which also reported that recurrent architectures tend to plateau with increasing temporal context, whereas Transformers can more effectively exploit longer input horizons.

3.4. Explaining models

To assess the general relevance of input features for wildfire danger prediction, we computed mean absolute SHAP values for each feature, averaged across all evaluated models. Figure 3(a) 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 *rh*, *2m temperature (t2m)*, and *total precipitation (tp)*. Similar prominence of daytime land surface temperature has been reported in recent ML-based wildfire prediction studies [24, 92, 93]. In contrast, 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 DL models (figure S3(a)) 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 [94]. 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 3(b) 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 and XGB differ markedly from DL models. Both 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. To assess the robustness of these attribution patterns with respect to model perturbations, we evaluated SHAP-based feature rankings across Optuna-tuned model configurations with different hyperparameter settings and across cross-validation folds. The rankings remained qualitatively consistent among models with similar predictive performance. For comparison, the corresponding IG-based feature ranks are shown figure S3(b). 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. In essence, data-driven models of the kind employed here are primarily aimed at maximizing predictive power rather than capturing actual process-based physical relationships or causality [95–97], also subject to the choice and availability of predictors (how the regression problem is framed). Still, as long as predictive accuracy generalizes across a large range of realistic settings, we argue that the resulting models have value and should be checked for basic compatibility with scientific intuition. Consequently, the xAI results presented here should not be overinterpreted but can serve to characterize DL model behavior, including broad consistency with scientific knowledge, and potential inconsistencies in behavior across DL architectures. 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 S5), which may cause attribution scores to be split or inconsistently distributed across redundant features [98, 99]. However, 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 purely 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. Moreover, we also note the intrinsic uncertainties in SHAP value approximation methods, especially in (close) relative variable rankings [100].

While our primary analysis focuses on marginal feature relevance, we note that variable effects may be non-additive and context-dependent. As an illustrative example, we provide second-order accumulated local effects interaction plots for selected driver pairs in figure S7. A systematic and comprehensive investigation of interaction effects across models is left for future work. Despite these limitations,

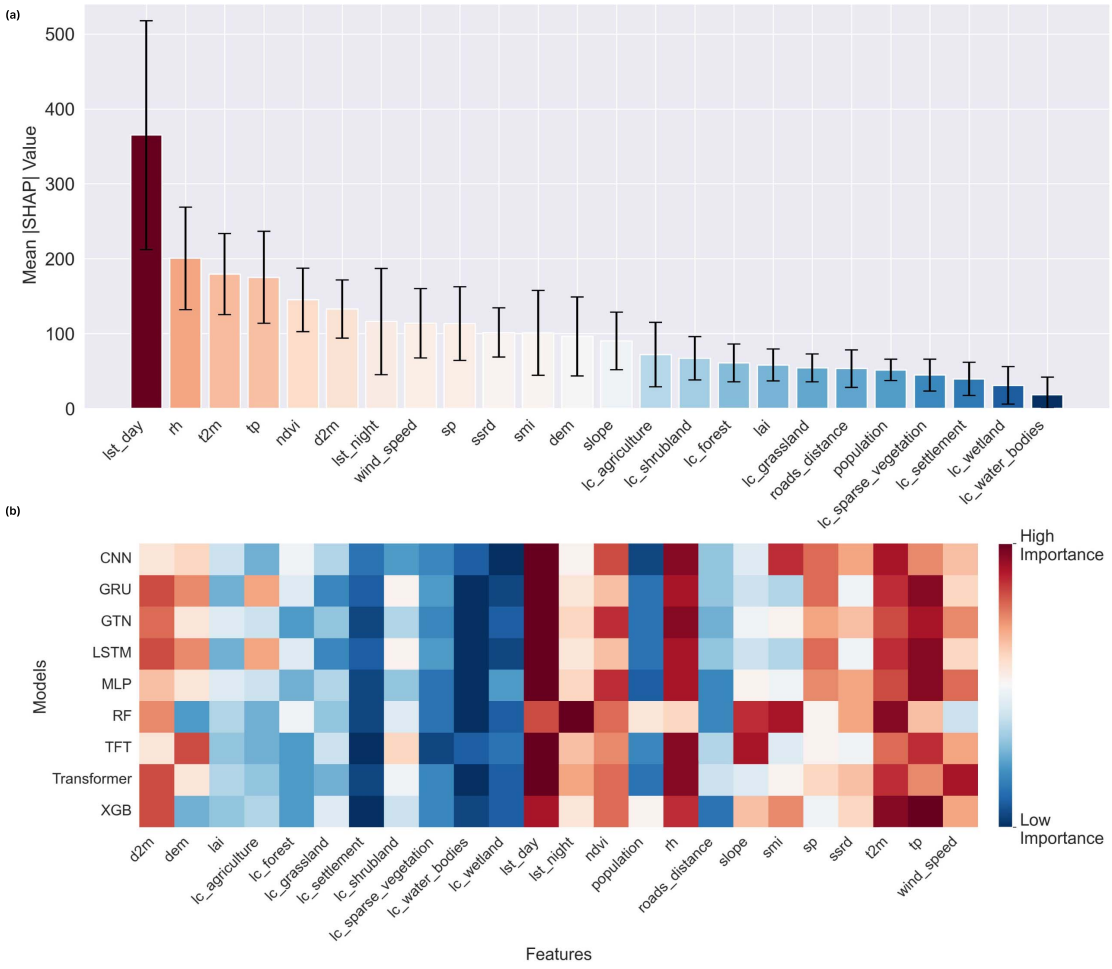

Figure 3. Cross-model SHAP-based feature analysis. (a): average absolute SHAP values per feature across all models, sorted in descending order of mean importance. Higher values indicate stronger influence on predictions. Bars show the mean absolute SHAP value, and error bars denote the standard deviation across models. (b): heatmap of relative feature ranks across models, with dark red indicating higher importance and blue lower relevance.

our analysis provides valuable insight into the models' prediction processes and allows us to evaluate whether the learned behavior is consistent with established scientific understanding of wildfire drivers, to the degree possible in a complex system of interacting meteorological, vegetation, and human factors.

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). 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 [101, 102], such as *ndvi*, *population*, *elevation (dem)*, *surface pressure (sp)*, and *LAI*.

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 4. 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* [103] and Di Giuseppe *et al* [21].

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 and the XGB baseline models, with the lowest F_1-scores, outperformed all others in terms of physical consistency, capturing 13 and 12 out of 19 relationships, respectively. This included several land cover classes representing different fuel availabilities (e.g. *lc_settlement*, *lc_waterbodies)* that were not correctly represented by any DL model. Certain

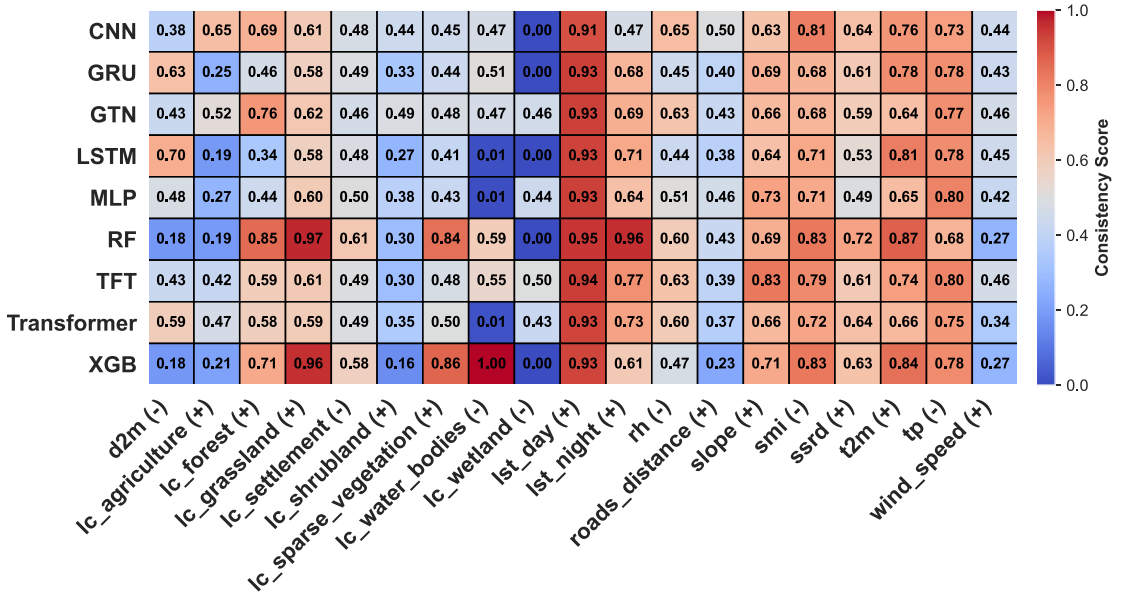

Figure 4. Physical consistency scores showing, for each feature-model pair, the proportion of samples in which SHAP-based attributions align with expected physical relationships from the fire science literature. The sign of the expected relationship with wildfire occurrence is indicated for each variable by ± in brackets. Darker colors highlight higher consistency within a model; the raw values are also provided as numbers.

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.

3.6. Case studies: comparison of two fire events in spain

Our xAI analyses so far focused on an average view on how much the various features contribute to test data predictions. However, these analyses did not reveal how the data-driven models arrived at individual predictions, which might predict fire or non-fire events for very different reasons, on a case-by-case basis. One might also wonder how sensitive these predictions are to specific input features, including the role and relative weighting of highly collinear variables. To address these questions and better understand the trade-offs between predictive skill and explainability, we examined two large fire events in Spain's Zamora province (Castilla y León) during summer 2022. We analyzed a false negative on June 15 and a TP on July 17 (figure S8). These case studies are indicated by star symbols in figure 2 (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 5 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*), (*rh*), and vegetation indices (*lai*, *ndvi*). These signals were less pronounced in the June case, which was reportedly ignited by lightning [104].

In figure 6(a), 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 rh 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 6(b). For the June fire, the predicted

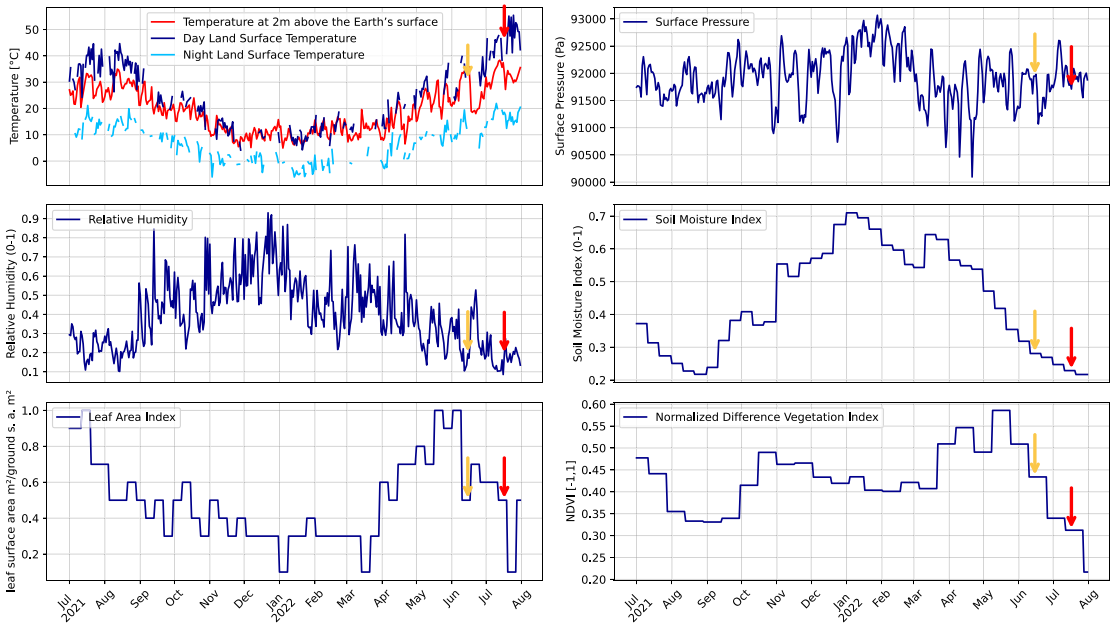

Figure 5. Temporal evolution of environmental variables before the two Spain fires 2022. The two arrows indicate the ignition dates: yellow for the June fire and red for the July fire. The data is shown at the average location between the two ignition points (41.82° N, 6.13° W). The two ignition points 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 d temporal context window prior to each event, rather than instantaneous values.

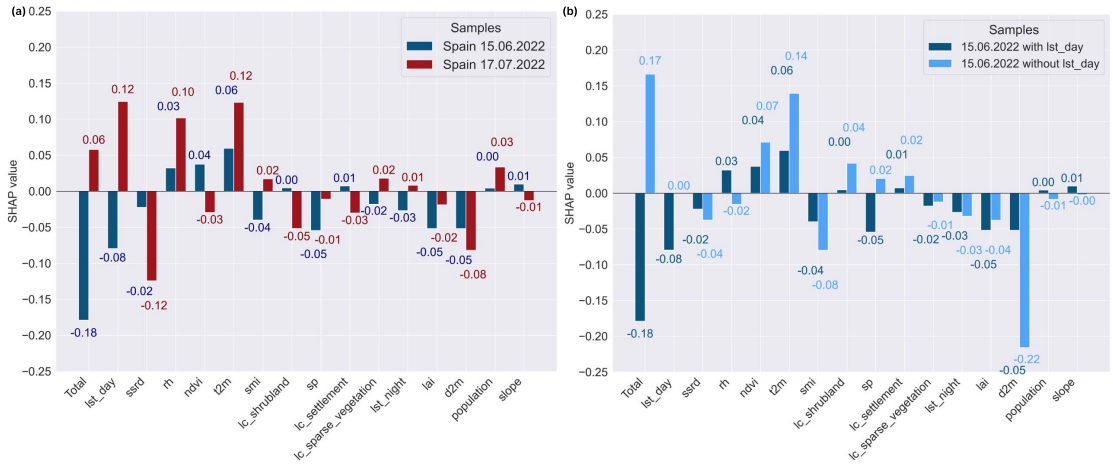

Figure 6. (a) SHAP attributions for the June fire (false negative in blue) and the July fire (true positive in red), highlighting differences in the influence of temperature- and humidity-related drivers. (b) SHAP attributions for the June fire with (dark blue) and without (light blue) the *lst_day* feature, showing that its exclusion increases fire probability and shifts contributions to other temperature variables. The absolute SHAP values are displayed above the bars.

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 S4). The IGs 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 rh 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.

4. Discussion & conclusion

Overall, all DL models achieved good predictive performance ($F_1 > 0.81$), outperforming the two baselines, consistent with Kondylatos *et al* [24], who also reported superior DL performance over RF for wildfire forecasting in Greece. Performance differences among DL models were not statistically significant at the 95% confidence level. Notably, more 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. Similar findings were reported by Kondylatos *et al* [24], where a 'vanilla' LSTM outperformed ConvLSTM, and by Di Giuseppe *et al* [21], who concluded that increased ML complexity does not inherently improve forecasts, and that data quality and representation of the fire triangle (i.e. weather, fuel and ignition) may be more influential than architectural sophistication [21].

Temporal analysis showed that Transformers leverage longer historical contexts more efficiently than LSTMs, with performance improving up to 30 d input windows. LSTM performance plateaued earlier, reflecting its limitation in capturing extended temporal dependencies. This corroborates with Prapas *et al* [31], who found that Transformers models degrade more gradually with increasing predictor dimensionality using the global Seasfire data cube [105]. Similarly, Michail *et al* [91] 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.

We observed training data imbalances (e.g. the proportion of coastal vs inland fire events), which were associated with reduced predictive performance in inland areas, reflecting a common challenge in ML [86–88]. Adopting a region-specific modeling strategy by subdividing the dataset by region, and training separate models for coastal and inland areas, as demonstrated in Schmitt *et al* [88], improved the F_1-score of inland fires by $\Delta 0.04$ by tailoring models to more homogeneous conditions. However, this comes at the expense of reduced generalization.

Across DL architectures, SHAP and IG showed largely consistent patterns of feature importance (figure 3(b)). The three most relevant predictors were temperature, rh, and precipitation, followed by NDVI and d2m. Although these results align with intuitive expectations and prior studies [24, 79], caution is warranted to avoid confirmation biases [106]. Moreover, human activity accounts for most wildfire ignitions [107–111], yet human-related variables were not identified as influential predictors. This likely reflects their coarse and indirect representation in the input data (e.g. country-level annual population density, distance to roads), rather than a limitation specific to the models themselves. The xAI analyses revealed a trade-off between predictive accuracy and physical plausibility. While the Transformer achieved the highest accuracy, the RF benchmark exhibited more physically consistent predictor-predictand relationships, including some that were not well captured by the DL models. The relative importance of these aspects depends on the intended application. In operational forecasting, predictive skill may be an important factor to reduce false negatives. In scientific contexts, physically interpretable relationships may be equally relevant, particularly when assessing potential drivers of wildfire risk under climate change scenarios [112–114]. While presenting explainability results alongside predictive performance does not imply that the model encodes human domain knowledge, it allows an assessment of whether learned model behavior is consistent with established physical understanding. Such analysis can support trust and transparency, help diagnose potential model biases, and provide guidance for future model refinement. A direction for future research lies in moving from descriptive to causal explainability, that is, from identifying which features drive model predictions, toward establishing why those features matter in terms of underlying physical mechanisms. Integrating causal inference frameworks, such as structural causal models, with post-hoc xAI methods would allow researchers to test whether the associations surfaced by model explanations reflect genuine causal pathways or merely statistical dependencies encoded in the training data. In this context, Mengaldo [115] proposes a framework in which interpretability outputs serve as a bridge between machine-learned representations and human domain knowledge, enabling hypothesis generation precisely in cases where model behavior diverges from established understanding.

Overall, our results show that temporal DL models, particularly Transformers, can significantly improve next-day wildfire danger forecasting compared to tree-based algorithms. Their advantage lies in their ability to leverage extended temporal context to enhance predictive skill. However, assessing physical plausibility remains important for interpreting model behavior and guiding model selection according to the intended application. Model performance is also influenced by data quality and representativeness. In particular, data imbalances and the coarse, indirect representation of certain drivers (e.g. human-related variables) may limit their spatial performance and the models' ability to reliably capture the physical relationships underlying wildfire risk. Future work should further investigate interactions among drivers and could adopt more causal perspectives. Systematic analyses of model behavior under different ignition mechanisms, for example human-induced versus naturally caused fires, could provide novel insights into fire mechanisms and support more robust fire prevention, control, and operational forecasting.

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).

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://orionlab.space.noa.gr/mesogeos/ [32]. The main scripts for model training, and analysis are archived in a publicly accessible repository https://doi.org/10.5281/zenodo.19605470, with documentation to facilitate replication of the results.

Assessing and Explaining Temporal Deep Learning Models for Wildfire Danger Prediction available at https://doi.org/10.1088/3049-4753/ae5aa0/data1.

Funding

P N was partially funded by the UK Natural Environment Research Council (NERC), Grant Number NE/V012045/1. Both authors were also partially supported by the Carl-Zeiss-Foundation through the project 'WOW—a World Model of Our World', for which PN serves as PI. Both authors were also partially supported by the Carl-Zeiss-Foundation through the project 'WOW—a World Model of Our World'.

Author contributions

Pauline Becker 0009-0009-6298-1258
Conceptualization (equal), Data curation (lead), Formal analysis (lead), Investigation (lead), Methodology (equal), Project administration (lead), Resources (lead), Software (lead), Validation (equal), Visualization (equal), Writing – original draft (lead), Writing – review & editing (equal)

Carolina Natel 0000-0003-3103-6789
Data curation (supporting), Formal analysis (supporting), Funding acquisition (equal), Investigation (supporting), Project administration (supporting), Supervision (supporting), Validation (equal), Visualization (supporting), Writing – original draft (supporting), Writing – review & editing (equal)

Peer Nowack 0000-0003-4588-7832
Conceptualization (equal), Funding acquisition (equal), Investigation (equal), Methodology (supporting), Project administration (equal), Supervision (lead), Visualization (supporting), Writing – original draft (supporting), Writing – review & editing (equal)

References

[1] Prapas I, Ahuja A, Kondylatos S, Karasante I, Panagiotou E, Alonso L, Davalas C, Michail D, Carvalhais N and Papoutsis I 2023 Deep learning for global wildfire forecasting (arXiv:2211.00534)
[2] Jones M W *et al* 2022 Global and regional trends and drivers of fire under climate change *Rev. Geophys.* **60** e2020RG000726
[3] Perkins O, Kasoar M, Voulgarakis A, Edwards T, Haas O and Millington J D A 2025 The spatial distribution and temporal drivers of changing global fire regimes: a coupled socio-ecological modeling approach *Earth's Future* **13** e2024EF004770
[4] Rosu I-A, Mourgela R-N, Kasoar M, Boleti E, Parrington M and Voulgarakis A 2025 Large-scale impacts of the 2023 Canadian wildfires on the Northern Hemisphere atmosphere *npj Clean Air* **1** 22
[5] Grillakis M, Voulgarakis A, Rovithakis A, Seiradakis K D, Koutroulis A, Field R D, Kasoar M, Papadopoulos A and Lazaridis M 2022 Climate drivers of global wildfire burned area *Environ. Res. Lett.* **17** 045021
[6] Zheng B *et al* 2023 Record-high CO2 emissions from boreal fires in 2021 *Science* **379** 912–7
[7] Migliavacca M *et al* 2013 Modeling biomass burning and related carbon emissions during the 21st century in Europe *J. Geophys. Res.: Biogeosci.* **118** 1732–47
[8] Jones M W *et al* 2024 Global rise in forest fire emissions linked to climate change in the extratropics *Science* **386** eadl5889
[9] Rovithakis A and Voulgarakis A 2024 Wildfire aerosols and their impact on weather: a case study of the August 2021 fires in Greece using the WRF-Chem model *Atmos. Sci. Lett.* **25** e1267
[10] Driscoll D A *et al* 2024 Biodiversity impacts of the 2019-2020 Australian megafires *Nature* **635** 898–905
[11] Jones B A 2017 Are we underestimating the economic costs of wildfire smoke? An investigation using the life satisfaction approach *J. For. Econ.* **27** 80–90
[12] Bladon K D, Emelko M B, Silins U and Stone M 2014 Wildfire and the future of water supply *Environ. Sci. Technol.* **48** 8936–43
[13] Rovithakis A, Grillakis M G, Seiradakis K D, Giannakopoulos C, Karali A, Field R, Lazaridis M and Voulgarakis A 2022 Future climate change impact on wildfire danger over the Mediterranean: the case of Greece *Environ. Res. Lett.* **17** 045022
[14] Lee J H 2021 Prediction of large-scale wildfires with the canopy stress index derived from soil moisture active passive *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **14** 2096–102
[15] Shmuel A, Lazebnik T, Glickman O, Heifetz E and Price C 2025 Global lightning-ignited wildfires prediction and climate change projections based on explainable machine learning models *Sci. Rep.* **15** 7898
[16] Finney M A 2005 The challenge of quantitative risk analysis for wildland fire *For. Ecol. Manage.* **211** 97–108
[17] Sample M *et al* 2022 Adaptation strategies and approaches for managing fire in a changing climate *Climate* **10** 58
[18] Kolström M *et al* 2011 Reviewing the science and implementation of climate change adaptation measures in European forestry *Forests* **2** 961–82
[19] Menor I O *et al* 2025 Integrated fire management as an adaptation and mitigation strategy to altered fire regimes *Commun. Earth Environ.* **6** 202
[20] Jain P, Coogan S C, Subramanian S G, Crowley M, Taylor S and Flannigan M D 2020 A review of machine learning applications in wildfire science and management *Environ. Rev.* **28** 478–505
[21] Di Giuseppe F, McNorton J, Lombardi A and Wetterhall F 2025 Global data-driven prediction of fire activity *Nat. Commun.* **16** 2918
[22] Xu Z *et al* 2025 Deep learning for wildfire risk prediction: integrating remote sensing and environmental data *ISPRS J. Photogramm. Remote Sens.* **227** 632–77
[23] Li F, Zhu Q, Yuan K, Ji F, Paul A, Lee P, Radeloff V C and Chen M 2024 Projecting large fires in the western US with a more trustworthy machine learning method (available at: https://essopenarchive.org/users/784670/articles/948622-projecting-large-fires-in-the-western-us-with-a-more-trustworthy-machine-learning-method?commit=eeeede5be0d97a799b937bd99275926bd9e8da43)
[24] Kondylatos S, Prapas I, Ronco M, Papoutsis I, Camps-Valls G, Piles M, Fernández-Torres M-A and Carvalhais N 2022 Wildfire danger prediction and understanding with deep learning *Geophys. Res. Lett.* **49** e2022GL099368
[25] Sengupta A, Dutta R and Chanda K 2025 Interpretable machine learning for regional wildfire risk assessment: a comparative study of tree-based models *Ecol. Inform.* **80** 102541
[26] Malashin I P, Masich I, Nelyub V, Borodulin A, Gantimurov A and Tynchenko V 2025 Assessing wildfire extents in siberian forests using machine learning *Sci. Rep.* **15** 32834
[27] Lee C, Choi E H, Han Y and Lee Y 2025 Year-round daily wildfire prediction and key factor analysis using machine learning: a case study of gangwon state, south korea *Sci. Rep.* **15** 29910
[28] Yang C, Yao P, Wang Q, Wang S, Xing D, Wang Y and Zhang J 2026 Xgboost-based susceptibility model exhibits high accuracy and robustness in plateau forest fire prediction *Forests* **17** 74
[29] Liu J, Wang Y, Lu Y, Zhao P, Wang S, Sun Y and Luo Y 2024 Application of remote sensing and explainable artificial intelligence (xai) for wildfire occurrence mapping in the mountainous region of southwest china *Remote Sens.* **16** 3602
[30] Karasante I, Alonso L, Prapas I, Ahuja A, Carvalhais N and Papoutsis I 2025 SeasFire cube—a multivariate dataset for global wildfire modeling *Sci. Data* **12** 368
[31] Prapas I, Bountos N I, Kondylatos S, Michail D, Camps-Valls G and Papoutsis I 2023 TeleViT: teleconnection-driven transformers improve subseasonal to seasonal wildfire forecasting (arXiv:2306.10940)
[32] Kondylatos S, Prapas I, Camps-Valls G and Papoutsis I 2023 Mesogeos: a multi-purpose dataset for data-driven wildfire modeling in the Mediterranean (arXiv:2306.05144)
[33] Cilli R, Filipponi F and Gualdi S 2022 Interpretable machine learning models for wildfire prediction in the mediterranean region *Nat. Hazards Earth Syst. Sci.* **22** 2677–95
[34] Abdollahi A and Pradhan B 2023 Explainable artificial intelligence (xai) models for wildfire susceptibility mapping *Geocarto Int.* **38** 2163839
[35] Liao X, Zhang W and Wang Y 2025 Explainable AI for wildfire susceptibility mapping using shap and multi-source geospatial data *Int. J. Appl. Earth Obs. Geoinf.* **128** 103786
[36] Li H, Vulova S, Rocha A D and Kleinschmit B 2024 Spatio-temporal feature attribution of european summer wildfires with explainable artificial intelligence (xai) *Sci. Total Environ.* **916** 170330
[37] Zakari R Y, Malik O A and Wee-Hong O 2025 Spatio-temporal wildfire forecasting in australia using deep learning and explainable AI *Model. Earth Syst. Environ.* **11** 425
[38] Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat P 2019 Deep learning and process understanding for data-driven Earth system science *Nature* **566** 195–204

[39] Muñoz-Sabater J et al 2021 ERA5-Land: a state-of-the-art global reanalysis dataset for land applications *Earth Syst. Sci. Data* **13** 4349–83
[40] Wan Z, Hook S and Hulley G 2015 MOD11A1 MODIS/Terra land surface temperature and the emissivity daily L3 global 1km SIN grid *NASA LP DAAC*
[41] Didan K 2015 MOD13A2 MODIS/Terra vegetation indices 16-day L3 global 1km SIN grid V006 (available at: https://lpdaac.usgs.gov/products/mod13a2v006/)
[42] Myneni R B, Shabanov N V, Knyazikhin Y, Yang W, Dong H, and Tan B 2002 *MOD15A2: Global LAI and FPAR* vol 2002 pp B61B–0719 (available at: https://ui.adsabs.harvard.edu/abs/2002AGUFM.B61B0719M)
[43] Cammalleri C, Vogt J V, Bisselink B and de Roo A 2017 Comparing soil moisture anomalies from multiple independent sources over different regions across the globe *Hydrol. Earth Syst. Sci.* **21** 6329–43
[44] Tatem A J 2017 WorldPop, open data for spatial demography *Sci. Data* **4** 170004
[45] Franks S and Rengarajan R 2023 Evaluation of copernicus DEM and comparison to the DEM used for landsat collection-2 processing *Remote Sens.* **15** 2509
[46] Potapov P et al 2022 The global 2000–2020 land cover and land use change dataset derived from the landsat archive: first results *Front. Remote Sens.* **3** 856903
[47] San-Miguel-Ayanz J, Barbosa P M, Schmuck G, Libertà G and Meyer-Roux J 2003 The European forest fire information system (EFFIS) *Proc. 6th AGILE Conf. on Geographic Information Science* pp 24–26
[48] Giglio L, Schroeder W and Justice C O 2016 The collection 6 MODIS active fire detection algorithm and fire products *Remote Sens. Environ.* **178** 31–41
[49] Huot F, Hu R L, Ihme M, Wang Q, Burge J, Lu T, Hickey J, Chen Y-F and Anderson J 2021 Deep learning models for predicting wildfires from historical remote-sensing data (arXiv:2010.07445)
[50] Zhang G, Wang M and Liu K 2019 Forest fire susceptibility modeling using a convolutional neural network for Yunnan province of China *Int. J. Disaster Risk Sci.* **10** 386–403
[51] Bergstra J and Bengio Y 2012 Random search for hyper-parameter optimization *J. Mach. Learn. Res.* **13** 281–305
[52] Liashchynskyi P and Liashchynskyi P Grid Search, random search, genetic algorithm: a big comparison for NAS (arXiv:1912.06059v1)
[53] Li F et al 2023 AttentionFire_v1.0: interpretable machine learning fire model for burned-area predictions over tropics *Geosci. Model Dev.* **16** 869–84
[54] Akiba T, Sano S, Yanase T, Ohta T and Koyama M 2019 Optuna: a next-generation hyperparameter optimization framework *Proc. 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining* (*KDD '19*), (Association for Computing Machinery) pp 2623–31
[55] Rosenblatt F 1958 The perceptron: a probabilistic model for information storage and organization in the brain *Psychol. Rev.* **65** 386–408
[56] Hochreiter S and Schmidhuber J 1997 Long short-term memory *Neural Comput.* **9** 1735–80
[57] Cho K, v. Merrienboer B, Bahdanau D and Bengio Y 2014 On the properties of neural machine translation: encoder-decoder approaches (arXiv:1409.1259)
[58] Krizhevsky A, Sutskever I and Hinton G E 2017 ImageNet classification with deep convolutional neural networks *Commun. ACM* **60** 84–90
[59] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, u. Kaiser U and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* vol 30 (Curran Associates, Inc.) (available at: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html)
[60] Lim B, î Arık S, Loeff N and Pfister T 2021 Temporal fusion transformers for interpretable multi-horizon time series forecasting *Int. J. Forecast.* **37** 1748–64
[61] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32
[62] Chen T and Guestrin C 2016 XGBoost: a scalable tree boosting system *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 785–94
[63] Harris L and Grzes M 2019 Comparing explanations between random forests and artificial neural networks *IEEE Int. Conf. on Systems, Man and Cybernetics (SMC)* pp 2978–85
[64] Friedman J H 2001 Greedy function approximation: a gradient boosting machine *Ann. Stat.* **29** 1189–232
[65] Lek S and Park Y S 2008 Multilayer perceptron *Encyclopedia of Ecology*, ed S E Jørgensen and B D Fath (Academic) pp 2455–62
[66] Cui Z, Chen W and Chen Y 2016 Multi-scale convolutional neural networks for time series classification (arXiv:1603.06995 [cs])
[67] Kim D-K and Kim K 2022 A convolutional transformer model for multivariate time series prediction *IEEE Access* **10** 101319–29
[68] Rumelhart D E, Hinton G E and Williams R J 1986 Learning representations by back-propagating errors *Nature* **323** 533–6
[69] Sasaki Y 2007 The truth of the F-measure vol 1 p 5 (available at: https://nicolasshu.com/assets/pdf/Sasaki_2007_The%20Truth%20of%20the%20F-measure.pdf)
[70] Raschka S 2014 An overview of general performance metrics of binary classifier systems (arXiv:1410.5330)
[71] Bradley A P 1997 The use of the area under the ROC curve in the evaluation of machine learning algorithms *Pattern Recognit.* **30** 1145–59
[72] Bommer P L, Kretschmer M, Hedström A, Bareeva D and Höhne M M-C 2024 Finding the right XAI method-A guide for the evaluation and ranking of explainable AI methods in climate science *Artif. Intell. Earth Syst.* **3** e230074
[73] Turbé H, Bjelogrlic M, Lovis C and Mengaldo G 2023 Evaluation of post-hoc interpretability methods in time-series classification *Nat. Mach. Intell.* **5** 250–60
[74] Weng X, Forster G L and Nowack P 2022 A machine learning approach to quantify meteorological drivers of ozone pollution in China from 2015 to 2019 *Atmos. Chem. Phys.* **22** 8385–402
[75] Kuhn-Régnier A, Voulgarakis A, Nowack P, Forkel M, Prentice I C and Harrison S P 2021 The importance of antecedent vegetation and drought conditions as global drivers of burnt area *Biogeosciences* **18** 3861–79
[76] Hickman S H M, Griffiths P T, Nowack P J and Archibald A T 2023 Short-term forecasting of ozone air pollution across Europe with transformers *Environ. Data Sci.* **2** e43
[77] Lundberg S M and Lee S-I 2017 A unified approach to interpreting model predictions *Advances in Neural Information Processing Systems* vol 30 (Curran Associates, Inc.) (available at: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html)
[78] Molnar C 2025 *Interpretable Machine Learning—A Guide for Making Black Box Models Explainable* 3rd edn (available at: https://christophm.github.io/interpretable-ml-book)

[79] Abdollahi A and Pradhan B 2023 Explainable artificial intelligence (XAI) for interpreting the contributing factors feed into the wildfire susceptibility prediction model *Sci. Total Environ.* **879** 163004
[80] Abdollahi A and Pradhan B 2021 Urban vegetation mapping from aerial imagery using explainable AI (XAI) *Sensors* **21** 4738
[81] Lundberg S M, Erion G, Chen H, DeGrave A, Prutkin J M, Nair B, Katz R, Himmelfarb J, Bansal N and Lee S-I 2020 From local explanations to global understanding with explainable AI for trees *Nat. Mach. Intell.* **2** 56–67
[82] Sundararajan M, Taly A and Yan Q 2017 Axiomatic attribution for deep networks *Int. Conf. on Machine Learning* (PMLR) pp 3319–28
[83] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z and Desmaison A 2017 Automatic differentiation in PyTorch
[84] Kruskal W H and Wallis W A 1952 Use of ranks in one-criterion variance analysis *J. Am. Stat. Assoc.* **47** 583–621
[85] Prapas I, Kondylatos S, Papoutsis I, Camps-Valls G, Ronco M, Fernández-Torres M-î, Guillem M P and Carvalhais N Deep learning methods for daily wildfire danger forecasting (arXiv:2111.02736 [cs])
[86] Ramachandra V 2025 Artificial intelligence in climate science: a state-of-the-art review (2020–2025)
[87] Castrejon D J, Wang C, Osmak D, Kukadiya B, Liu L, Giraldo M and Jiang X 2023 Machine learning-based california wildfire risk prediction and visualization *2023 Int. Conf. on Machine Learning and Applications (ICMLA)* pp 1212–7
[88] Schmitt E A, Zaremba E, Ananthavaram N, Liu L, Giraldo M and Jiang X 2024 Ecosystem-based wildfire risk prediction with machine learning *2024 IEEE Int. Conf. Big Data (BigData)* pp 7540–5
[89] Rosner B, Glynn R J and Lee M-L T 2006 The Wilcoxon signed rank test for paired comparisons of clustered data *Biometrics* **62** 185–92
[90] Pölz A, Blaschke A P, Komma J, Farnleitner A H and Derx J 2024 Transformer versus LSTM: a comparison of deep learning models for karst spring discharge forecasting *Water Resour. Res.* **60** e2022WR032602
[91] Michail D, Panagiotou L-I, Davalas C, Prapas I, Kondylatos S, Bountos N I, and Papoutsis I 2024 Seasonal fire prediction using spatio-temporal deep neural networks (arXiv:2404.06437)
[92] Dastour H and Hassan Q K 2024 A multidimensional machine learning framework for LST reconstruction and climate variable analysis in forest fire occurrence *Ecol. Inform.* **83** 102849
[93] Cheerala U B, Chirukuri V T, Gummadi V A K, Bhuyan J M and Damacharla P 2025 Probabilistic wildfire susceptibility from remote sensing using random forests and SHAP avialble at: https://ui.adsabs.harvard.edu/abs/2025arXiv251111680B
[94] Sturmfels P, Lundberg S and Lee S-I 2020 Visualizing the impact of feature attribution baselines *Distill* **5** e22
[95] Debeire K, Bock L, Nowack P, Runge J and Eyring V 2025 Constraining uncertainty in projected precipitation over land with causal discovery *Earth Syst. Dyn.* **16** 607–30
[96] Nowack P, Runge J, Eyring V and Haigh J D 2020 Causal networks for climate model evaluation and constrained projections *Nat. Commun.* **11** 1415
[97] Hickman S, Trajkovic I, Kaltenborn J, Pelletier F, Archibald A, Gurwicz Y, Nowack P, Rolnick D and Boussard J 2025 Causal climate emulation with Bayesian filtering (arXiv:2506.09891 [cs])
[98] Aas K, Jullum M and Løland A 2021 Explaining individual predictions when features are dependent: more accurate approximations to Shapley values *Artif. Intell.* **298** 103502
[99] Wilkinson S, Nowack P and Joshi M 2025 Process-based machine learning observationally constrains future regional warming projections *J. Geophys. Res.: Mach. Learn. Comput.* **2** e2025JH000698
[100] Huang X and Marques-Silva J 2024 On the failings of Shapley values for explainability *Int. J. Approx. Reason.* **171** 109112
[101] Bistinas I, Oom D, Sá A C L, Harrison S P, Prentice I C and Pereira J M C 2013 Relationships between human population density and burned area at continental and global scales *PLoS One* **8** e81188
[102] Li L-M, Song W-G, Ma J and Satoh K 2009 Artificial neural network approach for modeling the impact of population density and weather parameters on forest fire risk *Int. J. Wildland Fire* **18** 640–7
[103] Chuvieco E *et al* 2023 Towards an integrated approach to wildfire risk assessment: when, where, what and how may the landscapes Burn *Fire* **6** 215
[104] Copernicus Emergency Management Service data 2022 Devastating wildfire in Sierra de la Culebra, Spain—Copernicus (available at: www.copernicus.eu/en/media/image-day-gallery/devastating-wildfire-sierra-de-la-culebra-spain)
[105] Alonso L 2024 SeasFire cube: a global dataset for seasonal fire modeling in the earth system *Zenodo* (https://doi.org/10.5281/zenodo.13834057)
[106] Roscher R, Bohn B, Duarte M F and Garcke J 2020 Explainable machine learning for scientific insights and discoveries *IEEE Access* **8** 42200–16
[107] Mukunga T, Forkel M, Forrest M, Zotta R-M, Pande N, Schlaffer S and Dorigo W 2023 Effect of socioeconomic variables in predicting global fire ignition occurrence *Fire* **6** 197
[108] DeFries R S, Morton D C, van der Werf G R, Giglio L, Collatz G J, Randerson J T, Houghton R A, Kasibhatla P K and Shimabukuro Y 2008 Fire-related carbon emissions from land use transitions in southern Amazonia *Geophys. Res. Lett.* **35** 2008
[109] Bistinas I, Harrison S P, Prentice I C and Pereira J M C 2014 Causal relationships versus emergent patterns in the global controls of fire frequency *Biogeosciences* **11** 5087–101
[110] Knorr W, Kaminski T, Arneth A and Weber U 2014 Impact of human population density on fire frequency at the global scale *Biogeosciences* **11** 1085–102
[111] Andela N and van der Werf G R 2014 Recent trends in African fires driven by cropland expansion and El Niño to La Niña transition *Nat. Clim. Change* **4** 791–5
[112] Perr-Sauer J *et al* 2025 Applications of explainable artificial intelligence in renewable energy research *Energy Rep.* **14** 2217–35
[113] Vasconcelos R N, de Santana M M M, Costa D P, Duverger S G, Ferreira-Ferreira J, Oliveira M, Barbosa L d S, Cordeiro C L and Rocha W J S F 2025 Machine learning model reveals land use and Climate's role in Caatinga wildfires: present and future scenarios *Fire* **8** 8
[114] Bhattarai H, Martin M V, Sitch S, Yung D H Y and Tai A P K 2025 Global patterns and drivers of climate-driven fires in a warming world *EGUsphere* **2025** 1–28
[115] Mengaldo G 2024 Explain the black box for the sake of science: the scientific method in the era of generative artificial intelligence (arXiv:2406.10557)