Deep Learning Emulation of Multivariate Climate Indices: A Case Study of the Fire Weather Index in the Iberian Peninsula

Óscar Mirones[1], Joaquín Bedia[2,3], Pedro M.M. Soares[4], José M. Gutiérrez[1], and Jorge Baño-Medina[1,5]

[1]Instituto de Física de Cantabria (IFCA), CSIC-Universidad de Cantabria, Santander, Spain
[2]Dept. Matemática Aplicada y Ciencias de la Computación (MACC), Universidad de Cantabria, Santander, Spain
[3]Grupo de Meteorología y Computación, Universidad de Cantabria, Unidad Asociada al CSIC, Santander, Spain
[4]Instituto Dom Luiz (IDL) - Faculdade de Ciências da Universidade de Lisboa (FCUL), Campo Grande Edifício C8, Piso 3, 1749-016 Lisboa
[5]Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California San Diego, San Diego, CA, USA

Correspondence: Jorge Baño-Medina (bmedina@ifca.unican.es)

Abstract. The Fire Weather Index (FWI) is an essential multivariate climate index for assessing wildfire risk and the associated impacts of climate change, as it provides a quantitative measure of wildfire danger by integrating different critical near-surface fire-weather variables, namely air temperature, relative humidity, wind speed, and precipitation. FWI calculation depends on instantaneous data representing noon local standard times, which are often unavailable in many climate data repositories—particularly in climate projections. In these instances, a "Proxy" of actual FWI is often used, applying the same FWI formulation to daily aggregated values (mean, max, or min), despite known limitations in capturing extremes and temporal dynamics.

This study investigates the use of deep learning (DL) models to emulate the reference FWI over the Iberian Peninsula—a predominantly Mediterranean and fire-prone region—using only daily inputs. The emulators are trained and evaluated using ERA5-Land data, which, while not observational ground truth, provides a consistent and high-resolution dataset suitable for controlled inter-comparison. The focus is not on validating FWI against observations, but on assessing the ability of DL models to reproduce the reference FWI more accurately than traditional proxy approaches, using the same input data source.

Our results show substantial improvements in spatial accuracy, preservation of temporal sequences, and detection of extreme fire danger events when compared with the corresponding Proxy version. Furthermore, after evaluating different combinations of input variables for DL model training, we find that precipitation can be excluded without substantially affecting accuracy—especially at the upper end—an important insight given the challenges climate models face in representing precipitation. These findings highlight the potential of deep learning tools to enhance the usability of FWI in contexts where sub-daily data are unavailable, and set the stage for the emulation of other multivariate climate indices, which are vital for climate impact studies, spatial planning and management, and adaptation decision-making.

1 Introduction

Fire is a global phenomenon that has shaped ecosystems since the emergence of land vegetation, influencing vegetation structure, the carbon cycle, and climate (Bowman et al., 2009). From a human perspective, wildfires can cause severe ecological damage, threaten lives and infrastructure, and degrade air and water quality, with far-reaching social and economic consequences (Bowman et al., 2017; Pausas and Keeley, 2021).

Fire danger indices are essential tools for wildfire risk assessment and climate impact studies. By integrating key atmospheric variables—such as temperature, humidity, wind speed, and precipitation—into a single, physically interpretable indicator (Fugioka et al., 2009; Yu et al., 2023), these indices support fire management agencies in resource allocation and early warning systems (Di Giuseppe et al., 2020). They also help researchers understand how climate change alters fire-weather patterns (Bedia et al., 2015; Jolly et al., 2015; Bento et al., 2023; Santos et al., 2023; Gincheva et al., 2024) informing adaptation strategies and policy decisions in fire-prone regions (Costa et al., 2020; DaCamara et al., 2024).

Among the various fire-weather indices worldwide, the Canadian Fire Weather Index (FWI, van Wagner, 1987) is one of the most widely adopted globally. In Europe, it underpins the European Forest Fire Information System (EFFIS San-Miguel-Ayanz et al., 2013) and has been extensively used for the estimation of fire danger in vulnerable regions (Camia and Amatulli, 2009), such as Iberia (Santos et al., 2023), as well as for the assessment of future climate change scenarios in the Euro-Mediterranean region (Bedia et al., 2014; Dupuy et al., 2020) and the Iberian Peninsula (Bento et al., 2023).

Although the adoption of the FWI presents clear advantages for effectively characterizing fire danger weather situations, it also poses some hurdles from an implementation point of view. One significant challenge is that the FWI was originally designed for noon local standard time (LST) conditions, representing the highest fire danger time in mid-latitude regions. This means it requires instantaneous noon time input data of near-surface air temperature, relative humidity, 10-m wind speed and last 24-hour accumulated precipitation (Lawson and Armitage, 2008). However, such specific noon data are often missing from observational records. Furthermore, climate model databases quite often do not include sub-daily outputs due to the massive storage requirements (e.g. the Earth System Grid Federation –ESGF–, Cinquini et al., 2014), thus preventing the direct calculation of FWI from climate model simulations.

To circumvent the lack of instantaneous data, daily mean aggregated inputs have been commonly used for FWI estimation, in particular to obtain future projections from model simulations. However, the resulting daily mean FWI proxies cannot be reliably transformed to match their instantaneous counterparts, leading to inconsistent future projections (Herrera et al., 2013). Consequently, approximations are often employed to minimize FWI distortion. One of the pioneering studies in this area utilized different daily *proxy* variables for FWI calculation from climate model outputs, comparing various combinations of minimum relative humidity and maximum temperature as surrogates for noon-time outputs. This approach aimed to minimize the distortion of the climate change signal in the reference FWI, using simulations from a regional model for both historical and a moderate future warming scenario (Bedia et al., 2014). Since then, other studies have adopted proxies to produce more accurate FWI projections (Abatzoglou et al., 2018; Bento et al., 2023; Matteo et al., 2025), as a practical solution given the

unavailability of more precise data. Despite all these efforts, the approximate version of the FWI inevitably presents certain deviations from the original version that introduce added uncertainty (Bedia et al., 2014; Matteo et al., 2025).

Emulation offers a promising alternative in this context. Recent advancements in machine learning techniques, particularly deep learning (DL), across various fields suggest that these sophisticated nonlinear models have the potential to effectively represent and capture the intricate structures and nonlinear dynamics inherent in Earth systems (Duffy et al., 2023), including numerical model climate predictions (Rampal et al., 2024) and applications to downscaling (Baño-Medina et al., 2022; Bushenkova et al., 2024; Li et al., 2025; Johannsen et al., 2024; Soares et al., 2024), among others. In this new framework, the emulation function is determined by the neural network's coefficients, allowing for varying topologies to produce different plausible fields (Baño-Medina et al., 2024). This opens up a broad area of research where various techniques and architectures can be compared to evaluate their suitability for specific problems. It also introduces new challenges related to model interpretability, requiring methods of eXplainable Artificial Intelligence (XAI) that reveal how DL models make decisions. Such approaches play a key role in establishing the credibility of the outcomes among users (Dramsch et al., 2025) and for gaining valuable insights into the most influential variables and the physical interpretation of model's functioning (Yang et al., 2024; Baño-Medina et al., 2025). Among the variety of approaches available, saliency maps emerge as valuable tools aiding in the design and evaluation of DL climate downscaling approaches in general (González-Abad et al., 2023), and have been successfully used to unravel DL-based FWI predictions in previous studies (Mirones et al., 2025).

In this study, we explore the emulation of the Fire Weather Index (FWI) using DL models to approximate the index's behavior as closely as possible with commonly available input variables whose temporal frequency (daily) does not match the definition of the "reference FWI" (based on instantaneous noon local standard times). We build upon the reference FWI and the optimal Proxy FWI version presented in Bedia et al. (2014), which has been widely used in subsequent studies. We compare the emulation of the reference FWI using different combinations of input variables (hereafter predictor sets), including inadequate (daily mean), incomplete (removing some input variables), or suboptimal temporal frequency inputs (i.e. daily minimum relative humidity), which are more commonly available than the required instantaneous data. We test various DL model architectures and topologies and utilize explainable AI (XAI) techniques, such as saliency maps, to provide end users and practitioners with a solid understanding of the emulation process. Our results indicate that DL emulators outperform the traditional Proxy approach in all validation aspects, including spatial representation, preservation of the temporal sequence, and detection of extreme fire danger conditions. The daily mean inputs, which are most widely available in public repositories, are adequate for accurately emulating the reference FWI using the developed DL methods. Furthermore, our findings suggest that precipitation can be omitted from the predictor set without significantly compromising FWI representation accuracy. Overall, we demonstrate that by leveraging simpler inputs, DL emulators can enhance the accessibility and applicability of FWI in impact studies, facilitating a more efficient and widespread use in climate impact studies and decision-making.

2 Data and Methods

2.1 Input Data

All data used in this study were obtained from the ERA5-Land database (Muñoz-Sabater et al., 2021), distributed by the Climate Data Store (CDS) of the Copernicus Climate Change Service (Buontempo et al., 2022). ERA5-Land is a high-resolution reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). It offers detailed information on land surface variables, providing a consistent view of their evolution over several decades. ERA5-Land features a finer spatial resolution of approximately 9 km grid spacing, compared to ERA5 ($\sim$25 km Hersbach et al., 2020). It covers data from January 1950 to the present, with an hourly temporal resolution. The dataset includes various land surface parameters (such as soil moisture, soil temperature, snow cover, and surface runoff) to control the simulated land fields, ensuring the data remains accurate and consistent. ERA5-Land uses atmospheric variables from ERA5, like air temperature, wind speed and humidity, and presents hourly temporal resolution, thus providing the necessary input variables for the calculation of both reference and Proxy FWI versions, as well as the different predictor sets used in this study (Sec. 2.3).

2.2 Fire Weather Index calculation

The reference noon LST FWI is calculated from ERA5-Land daily records of four near-surface meteorological variables measured at 12 UTC ($\sim$ noon time in the Iberian Peninsula): 24-hour accumulated precipitation, instantaneous wind speed, relative humidity, and temperature. These inputs are processed via a set of empirical equations that yield six intermediate components characterizing fuel moisture dynamics across different fuel layers (van Wagner, 1987; Stocks et al., 1989), the influence of wind on fire spread, and the total fuel available for combustion. These components are then integrated to compute the FWI, a dimensionless index representing potential fire intensity under given meteorological conditions for a reference fuel type (mature pine stands).

A key complexity arises from the fact that certain FWI components, particularly those related to fuel moisture, exhibit different drying rates and retain memory of past conditions (Wotton, 2009). To address this, FWI initialization relies on default values for some components, leading to a spin-up period before stabilization. This period is generally brief, typically lasting a few days to weeks, especially during the fire season, when the effects of snow melting on soil moisture are minimal. Consequently, spin-up effects are deemed negligible due to the rapid stabilization of the FWI over time and any residual impact of spin-up is expected to be minimal in the final outputs (Bedia et al., 2018).

2.3 Predictors Sets

To emulate the reference FWI, we consider several predictor sets derived from ERA5-Land, summarized in Table 1. The initial experiments use P0, which builds on the same set of variables used in the optimal FWI approach from Bedia et al. (2014) (hereafter Proxy FWI). This allows us to assess whether the DL models can more accurately replicate the FWI transfer function when provided with the same predictors as the Proxy FWI. Accordingly, P0 includes 24-hour accumulated precipitation,

daily mean air temperature and wind speed, and minimum relative humidity. Proxy FWI is derived from the standard FWI formulation using the latter set of variables. The use of minimum relative humidity in P0 ensures the fairest possible comparison with Proxy FWI. However, in predictor sets P1 and P2, we intentionally replace minimum relative humidity with daily mean relative humidity. This is motivated by the fact that daily mean variables are more consistently provided by climate model outputs and reanalysis datasets, and we wanted to examine the performance of the emulator under such predictor-limited cases. Precipitation poses an additional challenge: it is one of the most difficult variables for climate models and reanalyses to represent reliably, due to its strong spatial and temporal variability, its dependence on complex physical processes, and the influence of multiple interacting factors. To examine whether robust performance can still be achieved in its absence, we exclude precipitation from predictor sets P1 and P2. Finally, in P2 we tried measuring the impact of using wind speed module versus wind speed zonal components, which could be relevant when determining fire danger. Other predictors sets have been tested, such as P1 using minumum relative humidity instead daily mean relative humidity or P0 including the 24-hour precipitation, without any added value compared to the configurations mentioned in the manuscript. These experiment results are shown in Figure D2.

Predictors sets		**Variable**					
	Description	**Temp.**	**Rel. Hum.**	**Precipitation**	**Wind speed**	**Eastward wind**	**Northward wind**
FWI	Reference FWI definition (van Wagner, 1987)	12UTC	12UTC	24-h accumulated (12UTC-12UTC)	12UTC	×	×
P0	"Best" Proxy FWI (Bedia et al., 2014)	DM	Min	24-h accumulated (12UTC-12UTC)	DM	×	×
P1	Daily Mean inputs without precipitation	DM	DM	×	DM	×	×
P2	Daily Mean inputs with wind components	DM	DM	×	DM	DM	DM

Table 1. Summary of the predictor sets assessed in this study. 12 UTC correspond to instantaneous model outputs verifying at that time. DM corresponds to Daily Mean values, and Max/Min to Daily Maximum/Minimum values. Precipitation is the daily (last 24 hours) accumulated value, from 12 to 12 UTC. Cells marked with × indicate that the variable is not included in the corresponding predictor set.

2.4 Deep Learning Methods

In this section, we briefly describe the DL architectures trained in this study. A visual summary of their topologies is presented in Figure 1. The selected architectures include the Fully Connected Dense (FCD) model, chosen as benchmark for its relative simplicity, as well as DeepESD and U-Net, which have been recently introduced in the literature for climate downscaling (Baño-Medina et al., 2022; González-Abad et al., 2023; González-Abad and Gutiérrez, 2024). Alternative architectures, such

as Convolutional Long Short-Term Memory (ConvLSTM), were also evaluated. However, owing to their inferior performance relative to the selected architectures and their greater computational demands, they are not presented in this manuscript. The DL models are tasked to learn a mapping between different sets of FWI-related input variables (Table 1) and the reference FWI over the Iberian Peninsula as output, by minimizing a loss function (Sec. 2.4.4). The model output is an emulated version of the reference ERA5-Land FWI at the same spatial resolution than its inputs. All models are trained with the same optimization parameters independently of the architecture, using the Adam optimizer with a learning rate set to 0.0001, a batch size of 64 and a maximum of 1,000 epochs. An early stopping mechanism is set to prevent overfitting, which stops the training process if there is no improvement in the validation loss after 30 consecutive epochs. Furthermore, the input data are standardized as part of the model preprocessing:

$$x_i^{'} = \frac{(x_i - \mu_i)}{\sigma_i} \tag{1}$$

where $x^{'}$ represents the standardized value, x represents the raw (unstandardized) value, and μ_i and σ_i represents the mean and standard deviation at gridpoint i. Parameters μ and σ have been computed relative to the training period 1979-2017 (per gridpoint).

2.4.1 Fully Connected Dense

The Fully Connected Dense (FCD) model is a DL architecture that relies exclusively on fully connected layers to process spatial input data. The input is first passed through two consecutive dense layers, each containing 50 neurons, with Rectified Linear Unit (ReLU) activation functions (Glorot and Bengio, 2010) applied to introduce non-linearity and enhance feature extraction. Following these hidden layers, the transformed feature representation is fed into a fully connected layer with 12,800 neurons, utilizing a linear activation function to produce the final output. This output is then reshaped to match the dimensions of the predictand, ensuring consistency with the target variable.

2.4.2 DeepESD

The DeepESD model (Baño-Medina et al., 2022) is designed as a combination of convolutional and dense layers envisaged to efficiently process spatial data. It consists of three convolutional layers with 50, 25, and 10 kernels, each using ReLU activation functions (Figure 1 for more details). After the final convolutional operation, the resulting feature maps are flattened into a one-dimensional vector, which is subsequently passed through a fully connected dense layer comprising 12,800 neurons, reshaping to match the dimensions of the predictand.

2.4.3 U-Net

The U-Net model adopts an encoder-decoder architecture designed for spatial data processing. The encoder progressively extracts hierarchical features through a series of Convolutional Blocks (ConvBlocks), each containing convolutional layers with

ReLU activation, followed by 2×2 max pooling operations that reduce spatial dimensions while increasing the number of feature maps. The encoder expands from 64 to 512 channels, while reducing the spatial resolution from (80,160) to (10,20). In the decoder, the model reconstructs the original resolution using 2×2 transposed convolutions (deconvolutions) that progressively upsample feature maps (Prince, 2023). This is followed by convolutional layers that refine the output. The final layer consists of a 1×1 convolution, followed by a linear activation function, producing an output of dimensions (1,80,160).

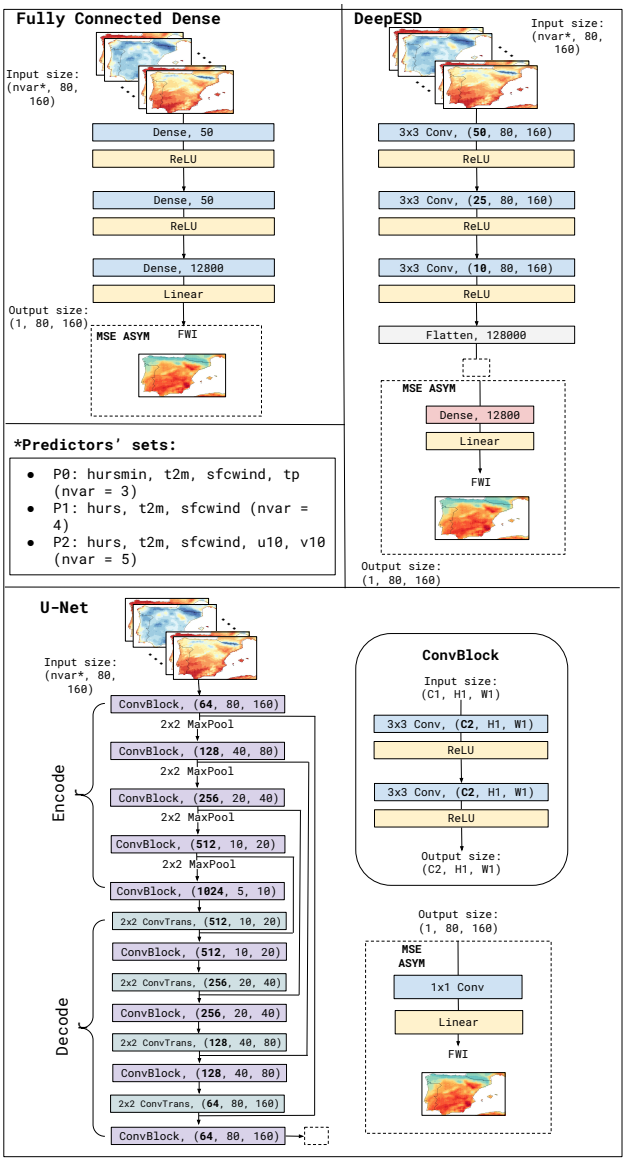

Figure 1. Schematic representation of the DL architectures used in this study, showcasing the different types of layers and their corresponding dimensions. The figure also highlights the distinct sets of predictors used as training inputs.

2.4.4 Loss functions

The DL models were fitted with two alternative loss functions: the mean squared error (MSE), which serves as a standard loss function, and an asymmetric loss function (ASYM), designed to better capture extreme events, as introduced by Doury et al. (2024). The MSE is defined as the average of the squared differences between observed and predicted values. This metric penalizes larger errors more severely, making it useful for many regression tasks. However, MSE might not adequately

emphasize errors associated with extreme events. To address this limitation, we implemented an asymmetric loss function defined as:

$$L_\theta = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| + \gamma^2 \times max(0, y_i - \hat{y}_i) \quad (2)$$

where $\gamma = G(y)$, where G is the gamma cumulative distribution function (CDF) fitted to the time series of every grid point in the training dataset. This loss function weights the mean absolute error (MAE) by an amount proportional to the extremity of each value, thus assigning a higher penalty when the model underestimates extreme values. This tailored approach improves the evaluation and prediction of extreme events in the dataset.

2.5 Model evaluation

In this study, the DL models are trained using daily data from the period 1979 to 2017. To evaluate the performance of the model, 20% of these data is separated and treated as validation samples for model assessment during the training phase. The final results are presented for an independent test period spanning 2018–2021. We also evaluated longer test periods (2012–2021), which required shortening the training phase to 1979–2011. Since these tests produced robust and consistent results comparable to those for 2018–2021, they are not included in the text. To assess the performance of the DL models and the reproducibility of spatial and temporal patterns in the emulated results, several validation indices are used, each focusing on different spatial and temporal aspects of the predicted series (Table 2).

Metric	**Description**
MAE FWI	Mean Absolute Error for the FWI
MAE FWI95	Mean Absolute Error for the FWI percentile 95-th
Freq. FWI95 Relative Bias	Relative bias of the frequency of extreme FWI events (over 95-th percentile)
Max Spell95 Bias	Inter-annual maximum spell length bias for extreme FWI events (over 95-th percentile)
ROC AUC	Receiver-operating characteristic (ROC) curve Area for extreme FWI events (over 95-th percentile)
R^2	Coefficient of determination

Table 2. Validation indices assessed in the study.

At each grid point, validation indices are computed, and their spatial distributions are depicted in the maps presented in various figures in Section 3. These indices span different key characteristics of the predicted series to assess how models emulate FWI, such as the mean FWI and extreme FWI distribution, or the length of spells for extreme FWI events. Instead of using MAE (for FWI and FWI95) or bias (for Max Spell95), we assess relative bias with respect to the frequency of FWI95, given by the following formula:

$$\text{Relative bias} = \frac{X - \hat{X}}{X} \times 100 \quad (3)$$

where X represents the observed and $\hat{X}$ the estimated value.

We also analyze the Receiver Operating Characteristic (ROC) curve, and the area Under the Curve (AUC), a graphical tool for evaluating the performance of binary classification models (Hogan and Mason, 2011), in this case applied to assess the DL model's ability to discriminate dangerous FWI events. It plots the true positive rate (sensitivity) against the false positive rate at various decision thresholds and quantifies the overall discriminative ability of the model, with values ranging from 0.5 (random classification) to 1.0 (perfect classification). A higher AUC indicates better model performance in distinguishing between positive and negative cases, making the ROC AUC particularly useful in unbalanced classification problems where traditional accuracy measures may be misleading. In particular, the classification evaluated in the ROC AUC analysis pertains to FWI events classified as 'extreme' (see Table 3), defined as those exceeding the 95th percentile of the reference FWI, according to the danger levels established by the Spanish Meteorological Agency (AEMET[1]), being thus impact-relevant for wildfire risk assessment in the Iberian Peninsula. In addition to the validation indices listed in Table 2, we also use scatter plots the assessment of the emulated FWI 95th percentile against reference FWI climatologies, where the R^2 coefficient of determination is used.

Table 3. FWI Classes According to AEMET (Based on Percentiles)

Level	FWI Percentile Range
Low	Below 40th percentile
Moderate	40th – 65th percentile
High	65th – 85th percentile
Very High	85th – 95th percentile
Extreme	Above 95th percentile

2.6 Model explainability

DL models have demonstrated remarkable predictive capabilities across various scientific fields, yet their complex architectures often limit interpretability. This challenge has motivated the growth of eXplainable Artificial Intelligence (XAI), which aims to provide tools for understanding the underlying relationships learned by machine learning models (McGovern et al., 2019). In this work, we apply saliency-based methods to investigate the internal logic behind our DL predictions. Specifically, we employ a variant of the Integrated Gradients (IG) technique (Sundararajan et al., 2017), which has been widely adopted in climate applications (González-Abad et al., 2023; Kondylatos et al., 2022) for its effectiveness and ease of interpretability.

Traditional IG computes feature attributions by integrating the gradients of the model output with respect to the input, along a linear path from a baseline input to the actual input. The method is formally expressed as:

$$s_i(x_i; x_b) = (x_i - x_b) \int_0^1 \frac{\partial f\left(x_b + t(x_i - x_b)\right)}{\partial x_i}\, dt \tag{4}$$

[1]https://www.aemet.es/documentos/es/datos_abiertos/Estadisticas/IM_riesgo_incendios/eimri_generalidades.pdf

where x_i is the feature value, x_b is the baseline (typically zero), and f is the output of the model.

However, in this study we adopt a simplified approximation of IGs, where the integral is omitted and the feature relevance is estimated directly from the gradients at the input point. This results in a computationally simpler, yet informative, saliency map representation. The relevance of each input feature is therefore computed as:

$$s_i(x_i) = x_i \cdot \frac{\partial f(x)}{\partial x_i} \tag{5}$$

This approach preserves the core idea of attributing the prediction to the input features based on the local sensitivity of the model, while avoiding the computational overhead of path integration. Although this approximation does not fully satisfy all axioms of the original IG method (Sundararajan et al., 2017), it has been shown to yield stable and interpretable attributions in practical applications (Toms et al., 2021).

To address known XAI challenges (Mamalakis et al., 2023), the raw saliency maps are post-processed by: (1) taking the absolute value to focus on magnitude only, (2) zeroing all values below 10% of the per-sample maximum to suppress gradient noise (Toms et al., 2021), and (3) normalizing each map so its values sum to one, thus yielding relative contributions. These normalized maps are used for all subsequent explainability analyses.

This methodology enables us to quantify the relative importance of each predictor variable in the model's decision-making process, providing valuable insights into the learned relationships and the adequacy of the different sets of predictors tested, aiding in the underlying physical interpretation of the results.

2.7 Software

Both FWI calculation and emulation are carried out using the R-based *climate4R* framework (Iturbide et al., 2019). In particular, neural network emulators (Dense, DeepESD, U-Net) are defined and trained using the *downscaleR.keras* package (Baño-Medina et al., 2022), an extension of the *downscaleR* package (Bedia et al., 2020) that leverages Keras/TensorFlow.

3 Results and Discussion

Several DL models are evaluated through an intercomparison of different predictor sets (Table 1) and the loss functions optimized during DL model training (Sec. 2.4.4). The results presented in this section focus on the main fire season over the Iberian Peninsula (June to September, JJAS; see e.g. Bedia et al., 2014) for the test period 2018-2021.

3.1 Proxy FWI validation

Figure 2 presents a multi-panel visualization, where the rows correspond to the validation measures shown in Table 2. From left to right, the first two columns represent the reference FWI and the Proxy FWI climatologies (P0, Table 1). The last column indicates the MAE or bias between both –depending on the evaluated index, Table 2–. The FWI95 frequency for the reference FWI is not shown, because, by construction, each grid point in the reference data records exactly 5% of days above its local

95th percentile threshold during the season. Therefore, the value in each grid across the spatial map is 0.05. It serves as the observed reference for error calculation: The 95th percentile of reference FWI is used as baseline to compute the frequency of Proxy FWI events exceeding this threshold (as well as emulated FWI in the following section).

Focusing on the differences between the Proxy and the reference FWI, we note that there exists greater MAE in Southern Portugal, North-western Spain and North-Eastern Spain, along the Ebro Valley. Conversely, the lowest MAE values (close to zero) are found along the Cantabrian and Mediterranean coasts, northwestern Portugal, the Pyrenees, and the Balearic Islands. These low-MAE areas generally align with the lowest climatological FWI values (as depicted in first column, first row), except for the Balearic Islands, where climatological JJAS FWI remains relatively high (FWI$\approx$35).

Regarding extreme FWI events (above reference FWI 95th percentile, FWI95), the Proxy FWI95 exhibits the highest MAE in the central-eastern and southeastern continental regions, coincident with very high reference FWI95 climatological values. Although the Proxy FWI95 accurately represents southern Portugal, revealing low MAE values, overall, the most critical FWI95 regions (mostly over Coastal Mediterranean and Central and Iberian Massifs) tend to be underrepresented by the Proxy FWI.

In regard with the temporal sequence of extreme event occurrence, the relative bias for the FWI95 frequency displays a spatial pattern similar to the Max Spell95, showing an overestimation of FWI95 events in Portugal, North-western Spain (Galicia) and also in the Cantabrian mountain range, the Pyrenees and the Ebro Valley, with a substantial underestimation in the rest of the Iberian peninsula. Regarding the spatial distribution of Max Spell95, the Proxy FWI overestimates spell lengths in Galicia and Portugal, with an excess of up to four days. Conversely, in the continental Iberian Peninsula, the Proxy FWI tends to underestimates spell length, remaining accurate over most of South-western Spain.

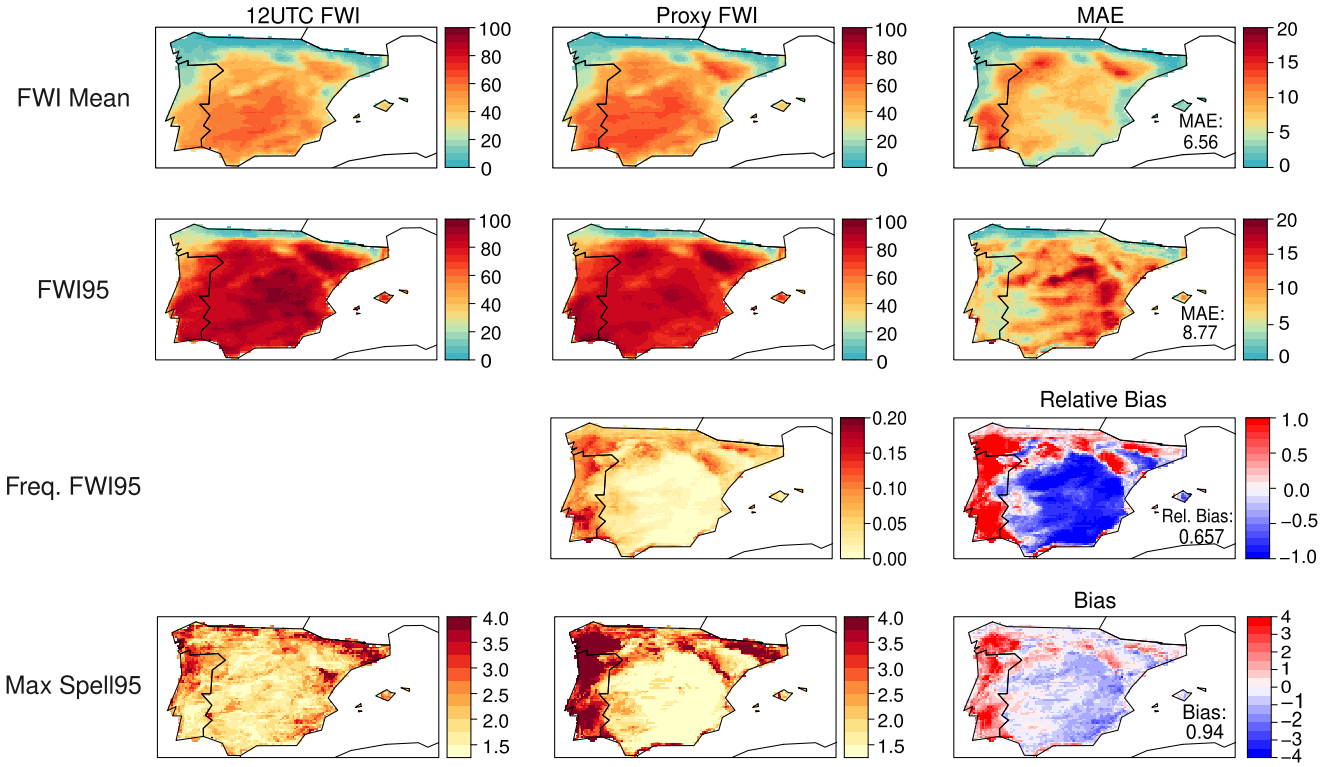

Figure 2. Reference and Proxy Fire Weather Index (FWI) climatologies for the fire season (June–September) during the test period (2018–2021). These include mean FWI, 95th percentile of FWI (FWI95), FWI95 frequency (number of days exceeding the reference FWI95) and the mean maximum annual spell (the number of days exceeding reference FWI95 Max Spell95). The third column displays the mean absolute error (MAE) for the FWI Mean and FWI95, the bias for Max Spell95 and the relative bias for the FWI95 frequency. The spatial averaged values of MAE or absolute (relative) bias are displayed at the bottom right.

3.2 Emulated FWI validation

This section presents an intercomparison of the different DL models tested, by evaluating errors relative to the reference FWI and comparing them with the Proxy FWI results described in the previous section, which serves as a benchmark. For brevity, here we present the results using the predictor set P0 as input (Table 1), considering that the overall intercomparison results are consistent regardless of the predictor set tested. Since P0 includes the variables necessary for computing the Proxy FWI using the standard FWI formulation, our goal is to assess whether different DL architectures are able to capture the reference FWI more accurately using the same input information.

3.2.1 Prediction errors

The prediction errors for the Dense, DeepESD, and U-Net models relative to the reference FWI are presented in Figure 3. The columns correspond to the results of each DL model, and the rows display various validation metrics, including the mean absolute error (MAE) for FWI and FWI95 or the relative bias in the frequency of FWI95. Below the error maps, a scatter plot is displayed per DL model, including the Proxy FWI, comparing the predicted climatological values against the reference FWI95.

Before discussing the performance of the DL models across the validation indices, we first highlight how daily aggregation of the input data affects model performance. Figure D1 in Appendix D presents the results from the Dense, DeepESD, and U-Net models trained using 12:00 UTC input variables (temperature, 24-hour accumulated precipitation, relative humidity, and wind speed) to compute the FWI. This sensitivity experiment evaluates the models' ability to learn the transfer function defining the FWI using inputs at 12:00 UTC, consistent with the temporal resolution of the reference index. This analysis complements our other model configurations by isolating the effect of temporal aggregation. Specifically, we compare performance when models are trained with instantaneous inputs (i.e., values at 12:00 UTC and 24-hour precipitation) versus daily mean inputs. This framework allows us to separate the intrinsic biases of each model from the additional error introduced by using daily-aggregated predictors. Figure D1 shows that models trained with instantaneous inputs exhibit lower bias and improved accuracy while maintaining a similar spatial error pattern. This confirms that part of the error observed in experiment P0 (Figure 3) stems from the mismatch in temporal resolution between the predictors and the reference FWI. However, some regions such as the Mediterranean areas for FWI MAE and the Cantabrian Mountains, the Pyrenees, and the Mediterranean coast for FWI MAE95 exhibit intrinsic errors even when instantaneous inputs are provided to the DL models. Moreover, the U-Net architecture demonstrates the lowest intrinsic bias in emulating the actual FWI function. It consistently shows the smallest biases in FWI and FWI95, as well as in the predicted frequency of FWI95 events and the mean annual maximum duration of FWI95 spells. These results suggest that U-Net offers enhanced generalization capabilities. Its ability to maintain low bias across both instantaneous and aggregated inputs indicates robustness to temporal variability, a key factor in modeling climate indices.

Focusing on the intercomparison of DL models trained with P0, all DL models exhibit similar performance across the validation indices compared to the Proxy FWI output. For each validation index, at least one DL model outperforms the Proxy FWI. While Dense reveals slightly worst mean FWI MAEs (Figure 3) than the Proxy FWI (Figure 2), DeepESD and U-Net has a comparable performance (6.54 *and* 6.01*vs.* 6.56). However, the DL models provide a smoother spatial distribution of MAE across both continental and coastal Mediterranean regions. In contrast, as discussed in Sec. 2.2, the Proxy FWI exhibits large MAE values in FWI-prone areas, such as the Ebro Valley, the Cantabrian mountain range, and southern Portugal (Figure 2), which are largely reduced by the DL emulated FWI (Figure 3).

Regarding FWI95 MAE, the U-Net model improves upon the Proxy FWI results (8.19 *vs.* 8.77), offering a smoother MAE distribution across the region and particularly low MAE values in some continental areas in Spain, where the Proxy FWI exhibits large errors. Except for Portugal, the Cantabrian and the Mediterranean coast, the U-Net model outperforms the Proxy

FWI across Iberia. Meanwhile, the other DL models display considerable MAE values, especially in central and northeastern Portugal and the Cantabrian mountain range, where these errors are larger that those of the FWI Proxy.

For FWI95 frequency, all three DL models outperform the Proxy FWI, with the U-Net model achieving the lowest relative bias. The Dense and DeepESD models generally underestimate FWI95 frequency across the region, except in southeastern Iberia and some localized areas in northern Iberia. Moreover, the U-Net model performs slightly better than the other DL models, overestimating FWI95 frequency across northern Iberia.

To assess the temporal characteristics of the emulated FWI, we evaluate Max Spell95 w.r.t. reference FWI. Here, the U-Net model achieves the best average result, closely followed by DeepESD (bias of 0.64 *vs.* 0.69). The key difference between these models is that U-Net generally overestimates is some northern areas, with maximum exceedances of 1–2 days, whereas DeepESD tends to overestimate, with discrepancies of up to 1-2 days in specific areas across the south-western of the Iberian Peninsula.

Despite these local problems, overall the scatter plots depict that DL models consistently improve upon the Proxy FWI results, regardless of the specific DL model. In terms of R^2, the U-Net model achieves the best performance for FWI95, with a value of 0.982, compared to 0.934 for the Proxy FWI. Overall, in the FWI95 scatter plots, the Proxy FWI underestimates higher values (above 60 for the reference FWI95), while the U-Net model, in particular, offers a significantly better representation of reference FWI.

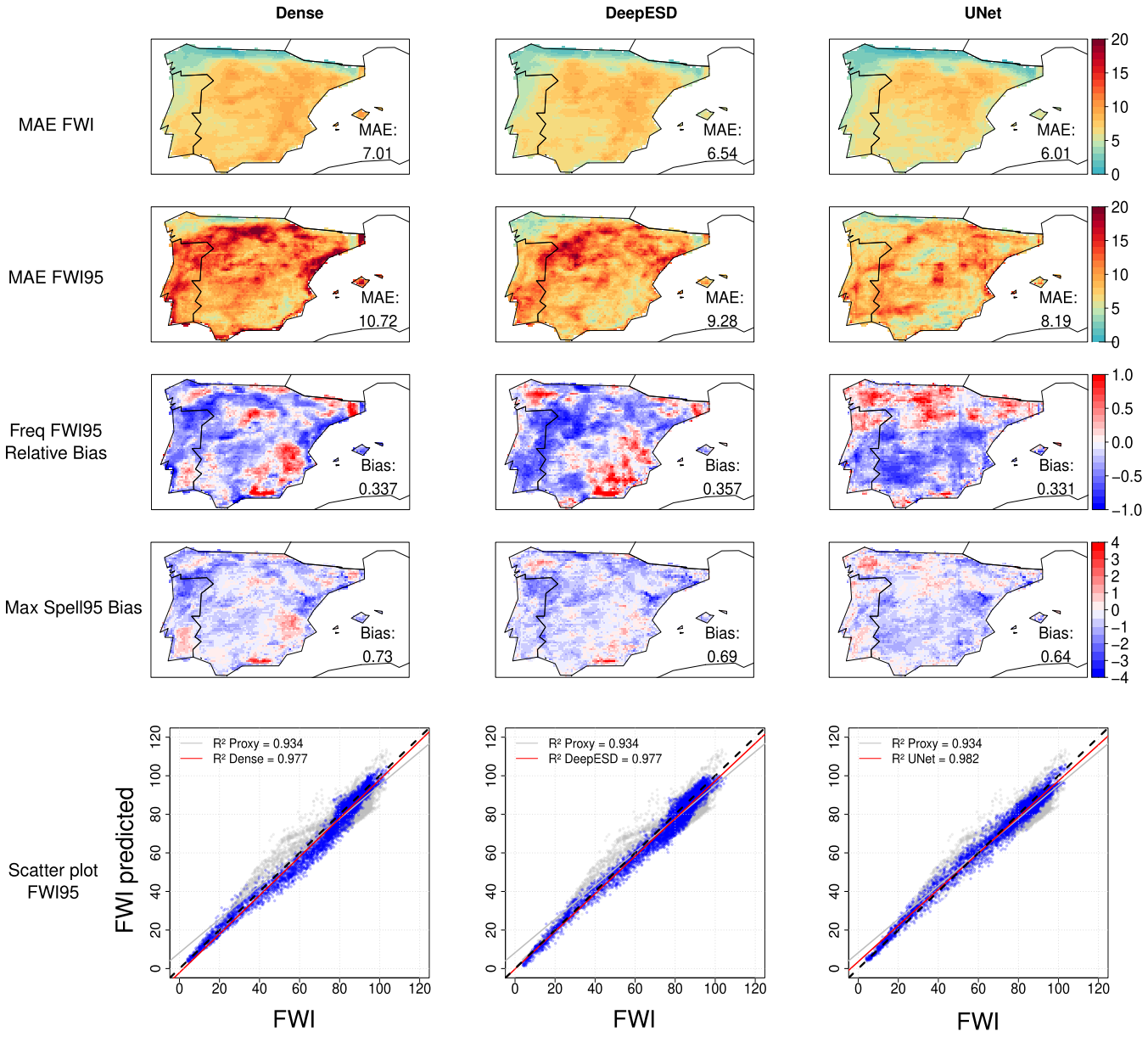

Figure 3. DL model results for various spatial validation indices presented in Figure 2. The results depict the differences relative to the reference FWI for the fire season (June–September) during the test period (2018–2021). The scatter plots compare climatology values between the Proxy FWI95 (in grey) and the corresponding DL model predictions (in blue), against reference FWI95. Linear fits are shown in grey for the Proxy FWI95 and in red for the DL model, with the corresponding R^2 values displayed in the top-left corner.

320 Furthermore in Figure 4, we assess the DL model ability in discriminating extreme fire danger events (i.e., reference FWI values above the 95th percentile) using ROC-AUC histograms, where the Proxy FWI is represented in red and the DL models in blue. Overall, the DL models improve upon the results provided by the Proxy FWI. Regarding extreme event detection,

DeepESD demonstrates the best performance, with a median ROC-AUC of 0.734, compared to 0.727 for U-Net and 0.7 for the Dense model, as shown in the histograms (Figure 4). In this regard, the DL models outperform the Proxy FWI, as indicated by the higher frequency of ROC-AUC values above 0.5 (indicating better-than-random classification). The only exception occurs for ROC-AUC values above 0.9, which correspond to the classification of grid points along the Cantabrian coast, where climatological FWI is the lowest Iberia (Figure 2) due to generally milder and moist conditions throughout the year, where Proxy FWI exhibits better FWI95 event discrimination than DL models.

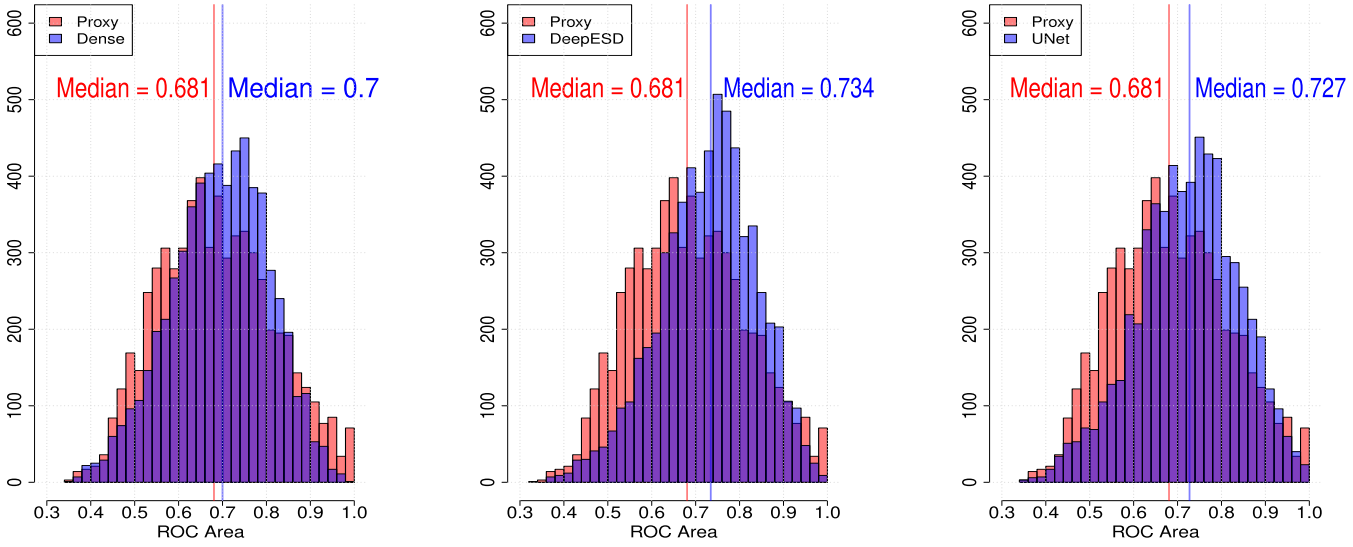

Figure 4. DL prediction results corresponding to the fire season (June–September) for the test period (2018–2021). Histograms compare the Area Under the ROC Curve (AUC)for extreme FWI event classification (above the 95th percentile) at each grid point, between the Proxy FWI (in red) and the corresponding DL model (in blue). The median values of both are also indicated.

In addition to the spatial assessment, we provide an analysis of the monthly boxplots of the Reference FWI, Proxy FWI, and the DL-predicted FWI (Dense, DeepESD, and U-Net) for each climatological region, as shown in Figure 5. Each boxplot represents the distribution of FWI values for a given month over the 2018–2021 period, capturing both the median and the spread of the values across the climatological regions (ATL, COASM, and CONTM; see Figure 2). This visualization demonstrates that the DL models not only reproduce the spatial patterns of FWI but also effectively capture the temporal variability across the year and across different regions. The median and interquartile ranges of the DL predictions closely follow the Reference FWI throughout the seasonal cycle, including the high-risk summer months, whereas the traditional Proxy FWI tends to underestimate extremes during peak months (June–September). The results further show that the Proxy FWI departs from the reference, particularly during the summer when values are highest and variability is greatest. In contrast, the DL models consistently produce distributions that are much closer to the Reference FWI, reducing the discrepancies observed with the proxy. This improvement is shown in all three regions, confirming that the advantage of the DL models over the Proxy FWI is

robust and not region-specific. The closer alignment of the DL predictions with the reference highlights their capacity to more accurately capture the seasonal dynamics of fire weather conditions without introducing significant seasonal biases.

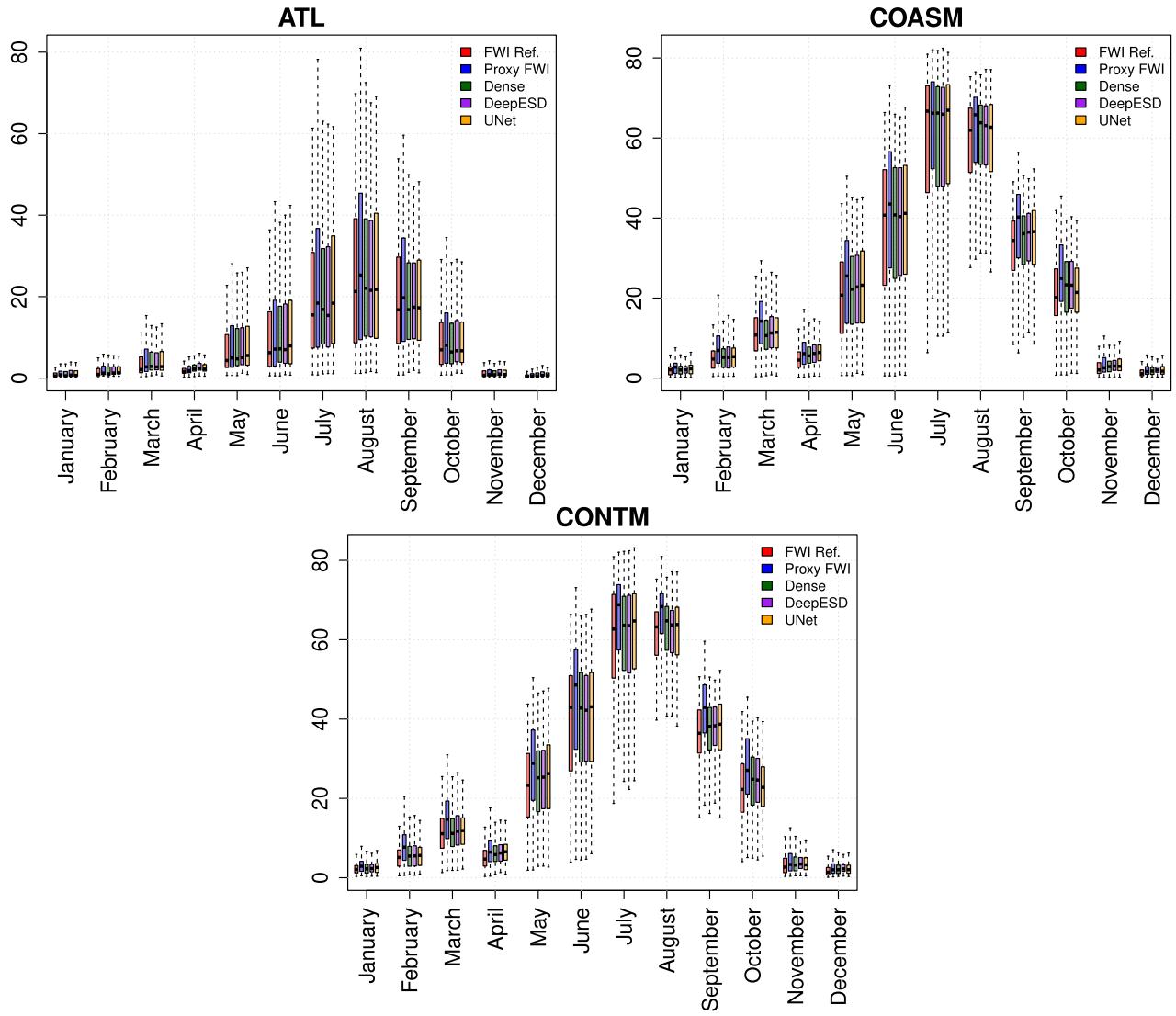

Figure 5. Monthly boxplots for the reference FWI, proxy FWI, and the DL-predicted FWI (Dense, DeepESD, UNet) for CONTM (top-left), COASM (top-right) and ATL (bottom) regions. Each boxplot represents the distribution of FWI values for a given month over the test period, capturing both the median and the spread of the values.

3.3 Emulation Model Explainability

Here, we analyze the saliency attributed to the predictor variables, grouped by the same climatic regions and fire danger categories defined in Appendix A. The overall results indicate no major differences in the saliencies of the different DL models,

underpinning their robustness and physical coherence. Therefore, here we present the explainability results of the U-Net model, since it outperforms Proxy FWI and attains similar validation results as DeepESD, while offering greater computational efficiency. To this aim, we compute the saliencies of each variable considering the time events corresponding to the different fire danger categories separately (Figure 6).

Focusing on the *high*, *very high*, and *extreme* fire danger categories, temperature and relative humidity emerge as the most influential variables in the model's predictions. The relative importance of these two variables varies across climatic regions. In the Continental Mediterranean region, temperature and relative humidity show similar levels of saliency. For the *high* fire danger category, relative humidity slightly surpasses temperature, whereas for the *very high* and *extreme* categories, temperature becomes more dominant. Wind speed consistently ranks as the third most important variable in these categories, although its relevance diminishes as the fire danger increases. Precipitation shows the lowest saliency in all categories and becomes increasingly negligible for higher danger categories.

This finding is consistent with the definition of the FWI itself, which relies primarily on temperature, relative humidity, and wind speed under dry conditions. The DL model thus reflects the structure and sensitivity of the FWI metric it is trained to emulate. Importantly, this does not mean that precipitation (or other inputs) is irrelevant for real-world fire risk; rather, it highlights that the predictand (FWI) gives limited weight to precipitation in high and extreme danger situations. Moreover, ERA5-Land predictor variables are not independent. Relative humidity, for instance, is derived from temperature and dew point, which are indirectly influenced by precipitation, while precipitation and wind are assimilated forcings. These interdependencies likely contribute to the model's ability to achieve high predictive accuracy even when precipitation and wind receive lower attribution scores.

The Coastal Mediterranean region shows a similar pattern. However, in the *high* category, temperature becomes more salient than relative humidity. In contrast, for the *extreme* category, relative humidity overtakes temperature as the most relevant predictor.

Similarly, in the Atlantic region, temperature is the most significant variable for the *high* and *very high* categories. For the *extreme* category, however, relative humidity surpasses temperature. Notably, in this region, the difference in saliency between wind speed and the top-ranked variables is smaller compared to other regions, suggesting a higher reliance on this variable in this region. As with the other regions, the relevance of wind speed decreases with increasing fire danger, and precipitation remains consistently negligible.

For the *medium* and *low* fire danger categories, the importance of variables shifts more markedly and shows less consistency across climatic regions. In the Continental and Coastal Mediterranean regions, precipitation becomes the most relevant variable for the *low* category, followed by relative humidity, wind speed, and temperature. For the *medium* category, wind speed emerges as the most salient predictor, followed by temperature, relative humidity, and precipitation.

In contrast, in the Atlantic region, relative humidity is the most important variable for the *low* category, followed by precipitation, wind speed, and temperature. For the *medium* category, temperature takes the lead, followed by wind speed, relative humidity, and finally precipitation. These results suggest that precipitation is a key variable for *low* fire danger levels but loses significance as fire danger increases, becoming practically irrelevant in *extreme* conditions, when temperature and relative hu-

midity gain prominence and become the most critical predictors. This is consistent with numerous studies that highlight the importance of high temperatures and low humidity in contributing to extreme fire danger (Jain et al., 2022).

Given that precipitation has negligible saliency in high percentile fire danger cases, which are precisely the most important from the point of view of impacts, the next section evaluates model performance after removing precipitation from the predictor set. We also examine alternative predictor sets, including replacements for precipitation and wind speed, such as eastward and northward wind components, under the hypothesis that these may carry additional useful information for the emulators.

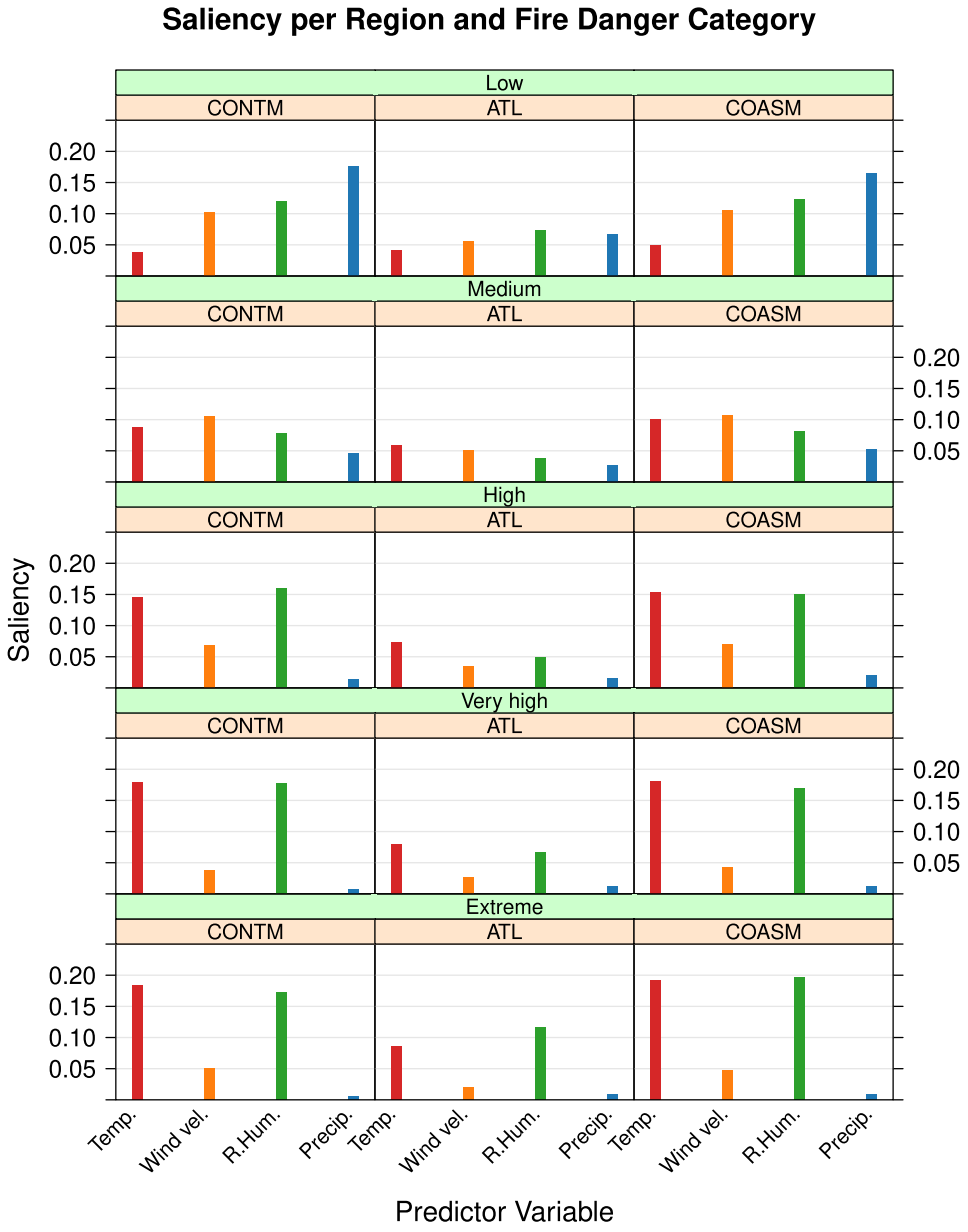

Figure 6. Aggregated saliency of predictor variables by climatic region and fire danger category for the U-Net model. Columns represent the climatic regions defined in Figure A1, while rows correspond to the fire danger categories described in Section 3.2. The displayed saliency values are derived from saliency maps, normalized at each grid point. Results for the JJAS season, test period (2018-2021).

3.4 Predictor set intercomparison

In this section, we intercompare different alternative input predictor sets, summarized in Table 1. For simplicity, we focus on the U-Net DL method as in Sec. 3.3. The U-Net trained with P0 and P1 yields better results than when trained with P2, except for the MAE FWI(Figure 7, analogous to Figure 3 for comparability). The difference between U-Net P0, U-Net P1 and U-Net P2 is relatively small, with the most remarkable variations observed in the western Continental region. For the MAE of FWI95 events, U-Net P2 significantly overestimates across the entire domain, whereas U-Net P0 and U-Net P1 present similar MAE spatial distribution, standing P1 as the best approach. However, U-Net P1 shows a greater presence of errors in specific regions, such as north Portugal, the Pyrenees and the Cantabrian montain ranges. Regarding the frequency of FWI95 events, U-Net P0 tends to overestimation across nearly the entire the north domain, while U-Net P1 tends to overestimate in the mid-south and east of the Peninsula and U-Net P2 exhibit a general underestimation. Among these, U-Net P0 appears to be the most effective in capturing this index. Examining the Max Spell95 results of the different experiments, U-Net P2 performs the worst, with overestimation up to 3 to 4 days in some areas, whereas the largest differences in P0 and P1 range between 2 and 3 days at most. U-Net P1 shows a slight improvement over U-Net P0, displaying more areas of overestimation, while U-Net P0 predominantly underestimates across most of the domain. Finally, the FWI95 scatter plots exhibit a high R^2 fit for all three predictor sets, outperforming the Proxy FWI. In terms of overall performance, U-Net P1 produces the best results, followed by U-Net P1 and U-Net P2, though the differences among the latter remain relatively small.

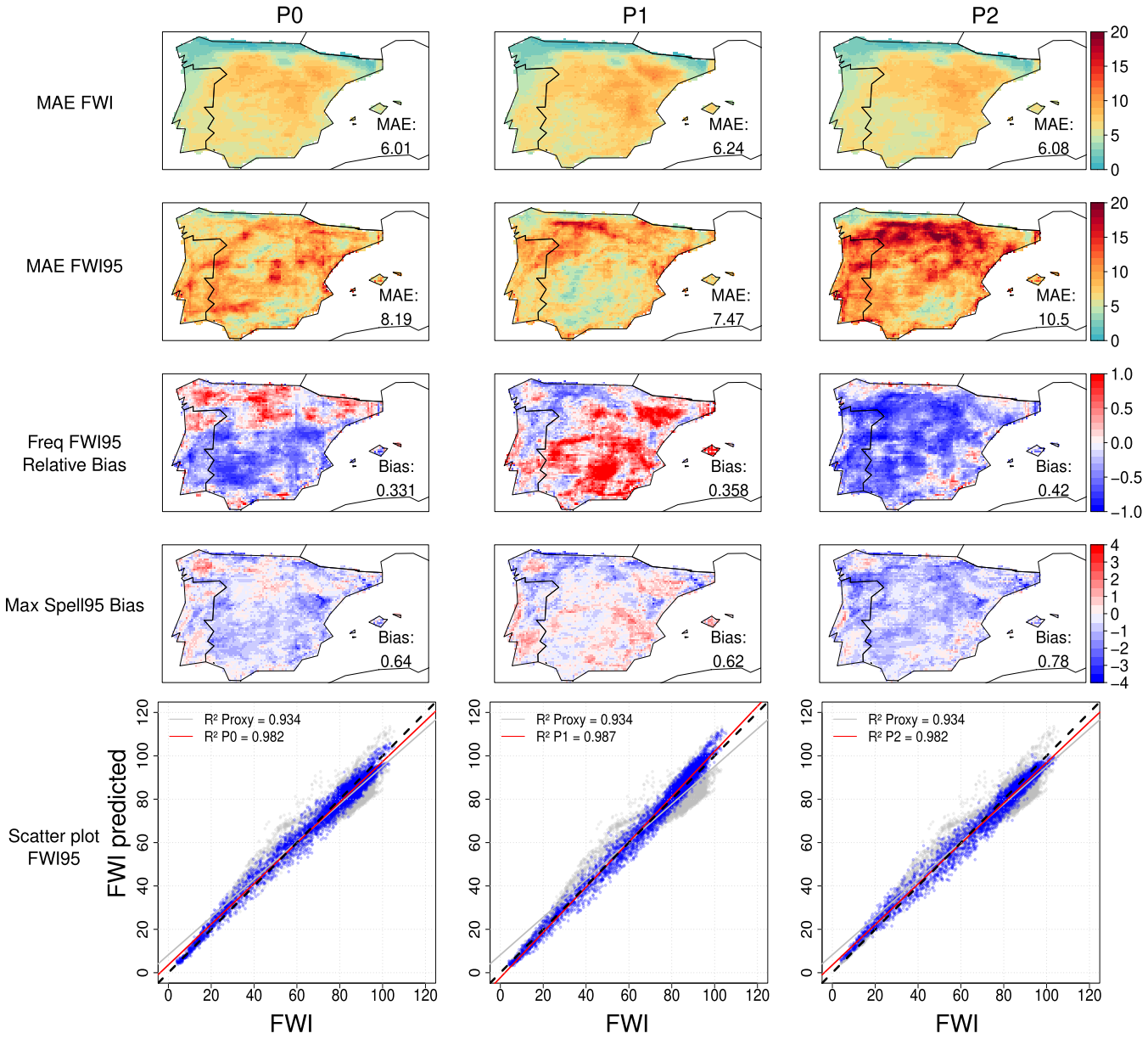

Figure 7. U-Net model results for various spatial validation indices presented in Figure 2 and different predictors sets assessed (see Table 1). The results depict the differences relative to reference FWI for the fire season (June–September) during the test period (2018–2021). Additionally, scatter plots compare climatology values of the Proxy FWI95 (in grey) and the U-Net model (in blue) against reference FWI95. Linear regression fit lines are shown in grey for the Proxy FWI95 and in red for the U-Net model (to enhance visualization), and the corresponding R^2 values displayed in the top-left corner.

Overall, in Figure 8, U-Net P1 yields the best results as classifiers for extreme events. These experiments improve upon the median results of the Proxy FWI, with U-Net P1 emerging as the most effective model in this context.

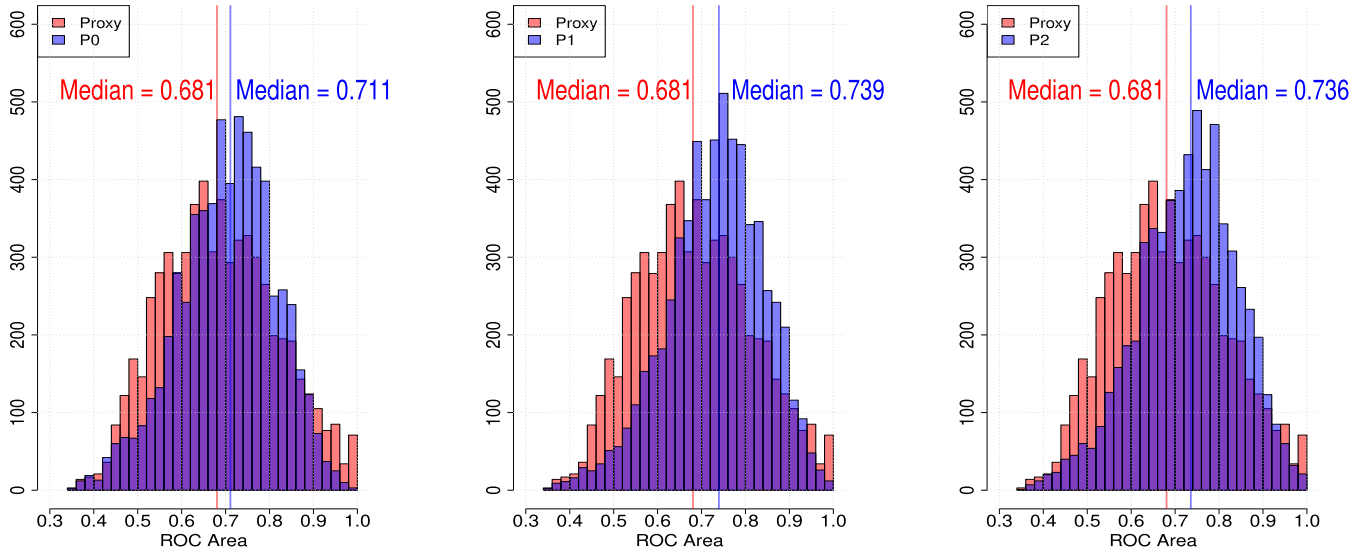

Figure 8. U-Net model histograms resultus for the fire season (June–September) for the test period (2018–2021). Histograms compare the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve for extreme fire danger classification (above the 95th percentile) at each grid point between the Proxy FWI (in red) and the U-Net model (in blue). The median value is indicated for both results.

In summary, the U-Net model demonstrates superior performance over the Proxy FWI, with U-Net P1 generally excelling in overall performance. The inclusion of wind components as additional predictors in P2 has a deleterious effect on the temporal and distributional similarity of emulated and reference FWI, suggesting that despite the selective use of features internally undergone by DL methods, a previous predictor screening might be necessary to ensure optimal results.

3.5 Loss function assessment

We next present a summary of the validation results, including an intercomparison of alternative MSE and asymmetric loss functions in the DL models (Sec. 2.4.4). Table 4 presents the validation indices for different predictor sets and DL models, comparing their performance against the Proxy FWI baseline, where the ASYM loss function values are indicated in parentheses. For all indices, the DL models exhibit consistent improvements over the Proxy FWI, with at least one model outperforming it in each category.

Regarding the loss function, the ASYM loss generally improves performance in most validation indices. The ASYM-trained models tend to yield lower MAE than ASYM values, particularly for DeepESD in P1 (6.74 vs. 6.43) and P2 (6.24 vs. 7.12), showing better accuracy in estimating FWI. Similarly, in the FWI95 MAE, the ASYM-trained models consistently outperform MSE loss counterparts, as seen with DeepESD in P0 (9.28 vs. 6.82) and P2 (9.66 vs. 9.44). In terms of extreme event prediction, the Max Spell95 Bias shows inconclusive results. While the ASYM loss function improves the bias in some cases, such as U-Net in P2 (0.78 vs. 0.54), in other cases, it worsens bias, as observed with Dense in P2 (0.78 vs. 1.05). This suggests that ASYM does not consistently improve the temporal persistence of extreme events but may still provide benefits in specific cases. For

the Relative Bias Frequency of FWI95, ASYM provides substantial benefits in reducing bias. The most notable case is U-Net in P1, where the ASYM-trained model achieves a significantly lower bias (0.358 vs. 0.269) than MSE, reinforcing ASYM's ability to enhance the reliability of extreme event frequency estimations.

Overall, ASYM appears to be particularly beneficial in reducing error in FWI estimations (MAE and MAE FWI95) and in improving the classification of extreme fire weather danger, as evidenced by the lower relative bias for FWI95 than MSE. However, its benefits on temporal aspects, such as Max Spell95 Bias, are not consistent. These findings suggest that while ASYM is generally advantageous, its effects are model- and index-dependent, although sometimes overestimates low and moderate values in the distribution (see Figure C1). Therefore, the ASYM loss function does not compromise the U-Net model performance in the lower quantiles, however for the Dense and DeepESD models low and moderate values for the FWI are worse than using the MSE as loss function. Therefore, the ASYM loss function can compromise the performance of low and moderate levels in some cases, and its selection should be carefully considered depending on the primary objective of the analysis.

Predictors sets	**DL Model**	**MAE**	**MAE FWI95**	**Max. Spell95 Bias**	**Rel. Bias Freq. FWI95**
Proxy FWI		6.56	8.77	0.94	0.657
P0	Dense	7.01 (7.94)	10.72 (8.86)	0.73 (0.95)	0.337 (0.662)
	DeepESD	6.54 (6.81)	9.28 (**6.82**)	0.69 (0.6)	0.357 (0.401)
	U-Net	**6.01** (6.09)	8.19 (8.51)	**0.64** (0.65)	0.331 (**0.313**)
P1	Dense	6.85 (7.33)	10.97 (9.83)	0.7 (0.92)	0.342 (0.514)
	DeepESD	6.74 (6.43)	8.43 (8.01)	1 (0.74)	0.666 (0.46)
	U-Net	6.24 (**5.84**)	**7.47** (8.46)	**0.62** (0.63)	0.358 (**0.269**)
P2	Dense	6.95 (7.58)	10.72 (9.8)	0.78 (1.05)	0.408 (0.588)
	DeepESD	6.24 (7.12)	9.66 (9.44)	0.69 (0.95)	0.362 (0.521)
	U-Net	**6.08** (6.95)	10.5 (**8.7**)	0.78 (**0.54**)	0.42 (**0.295**)

Table 4. Validation results of DL FWI emulators for different predictor sets. Values in parentheses are for the ASYM loss function for model training (default values correspond to MSE loss). The first row presents the Proxy FWI results, as benchmark. Bold values highlight the best validation index within each predictor set, while underlined values denote the overall best across all indices.

4 Conclusions

In this study, we have shown that DL emulators can effectively represent actual FWI using input variables of different temporal resolution and/or daily aggregation function as input. We have shown this for different DL methods and different sets of inputs using readily available ERA5-Land reanalysis predictors. Our analysis indicates that DL models not only capture the complex spatial and temporal characteristics of the reference FWI with higher fidelity than traditional Proxy-based approaches, but also improve the detection of most dangerous situations allowing for a more accurate classification of extreme fire danger

events. Our results strongly suggest that employing daily mean inputs, as opposed to instantaneous data, ensures adequate accuracy for operational FWI emulation. This finding is particularly valuable because daily datasets are more commonly accessible in climate repositories, thus lowering the barrier for integrating FWI predictions into fire management and climate risk assessments.

Furthermore, our investigation of different predictor sets revealed that excluding certain variables, such as precipitation, does not significantly degrade model performance. This suggests that simpler input configurations can be leveraged to achieve robust predictions, thereby reducing computational overhead and simplifying model implementation.

The XAI analysis reveals that temperature and relative humidity are the most influential predictors in high fire danger conditions, with their relative importance varying by climate regions. Wind speed holds moderate relevance but diminishes as fire danger intensifies, while precipitation consistently ranks as the least important variable and becomes nearly irrelevant under extreme fire danger conditions. In low and medium fire danger categories, the importance of variables has a greater regional variability, with precipitation emerging as a key factor in some cases. These findings underscore the critical role of temperature and relative humidity in high-to-extreme fire danger predictions and justify the consideration of alternative predictor sets that exclude precipitation as a non-essential predictor for dangerous event prediction.

The implications of our work extend beyond the immediate context of FWI estimation, and sets the stage for the emulation of multivariate climate indices relying on meteorological variables that are not commonly found in observational records or difficult to obtain from most climate model output repositories, given a sufficiently comprehensive database for its training (e.g., a high resolution reanalysis). By showcasing that DL models can serve as efficient surrogates for complex, computationally intensive fire danger indices, our study opens up new avenues for integrating AI into operational forecasting systems. This approach can significantly expedite the decision-making process in fire-prone regions, where timely and accurate risk assessments are critical.

Looking ahead, future research should focus on refining these models further by incorporating additional meteorological and environmental predictors, and by testing their transferability across different climatic regions. Ultimately, the continued development and operationalization of DL-based emulators have the potential to transform wildfire risk management, offering more precise, interpretable, and accessible tools for both researchers and practitioners.

Code and data availability. An illustrative example providing reproducible code and data via jupyter-notebook is provided in the open GitHub repository https://github.com/SantanderMetGroup/DeepFWI, where access to the required open data curated in Zenodo is granted and software environment configuration is detailed. Please note that due to brevity and the significant computing infrastructure required for full analysis reproducibility, the examples provided are a small sample of function calls and model configurations. These examples are applied on a coarser-than-native ERA5-Land grid for a limited time period, making them suitable for running locally on a CPU and avoiding extensive computing times. Further details or complete training/test datasets are available upon request to the authors.

Appendix A: Climatological regions

Here, we present the division of the Iberian Peninsula into three regions—Atlantic (ATL), Continental Mediterranean (CONTM), and Coastal Mediterranean (COASM)—for the eXplainable Artificial Intelligence (XAI) analysis provided in the main manuscript (Figure 6).

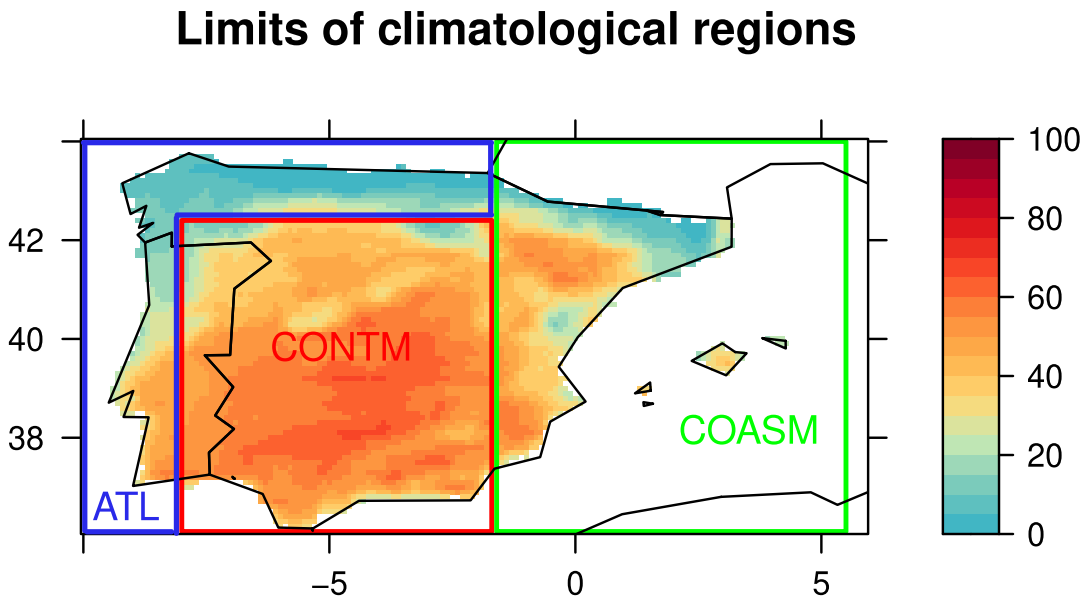

Figure A1. Climatological regions within the Iberian Peninsula used for XAI analysis in Figure 6. Blue, green and red rectangles indicate ATL (Atlantic), COASM (Coastal Mediterranean), and CONTM (Continental Mediterranean) regions respectively.

Appendix B: ERA5-Land limitations

In this section, we highlight the existent limitations in using ERA5-Land data in our analysis due to the inherited biases in ERA5-Land compared with observation data. ERA5-Land, like other reanalysis products, is subject to inherent biases that arise from the limitations of the numerical models and data assimilation techniques used in its generation. These biases reflect systematic deviations from ground-based observations and can affect the reliability of the dataset for certain applications. Therefore, although ERA5-Land provides a valuable, spatially and temporally consistent climate dataset, its outputs should be used with caution and validated against local observations whenever possible.

In Figure B1, we illustrate the ERA5-Land biases in some stations in Spain with respect to observation data provided by the Spanish Agency of Meteorology (AEMET).

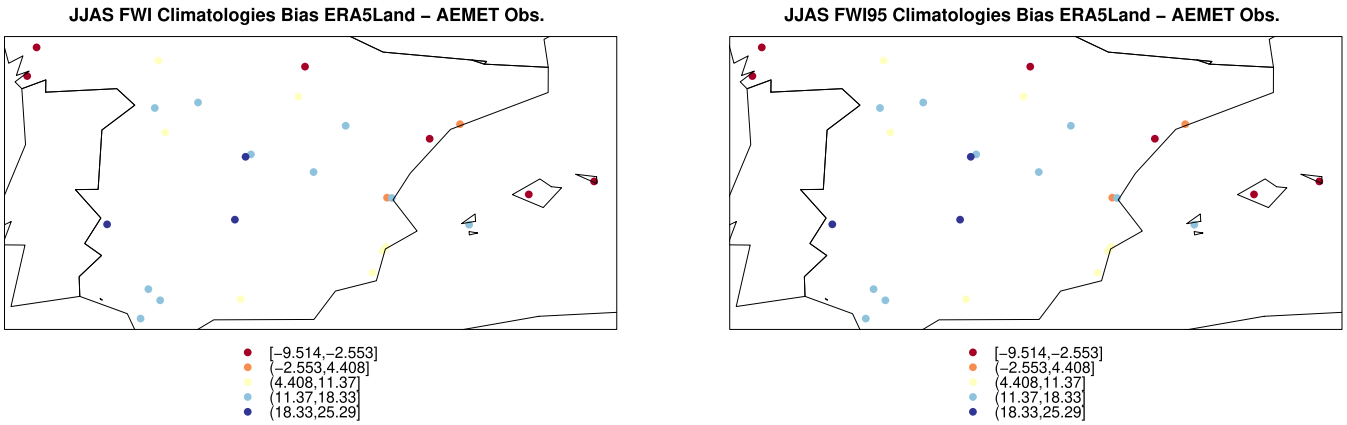

Figure B1. Biases between observational FWI data from the Spanish Agency of Meteorology (AEMET) and the FWI resulting from ERA5-Land computations. Top figure indicates de bias for the mean FWI, while bottom figure indicates it for the FWI95.

Appendix C: QQ-plot assessment

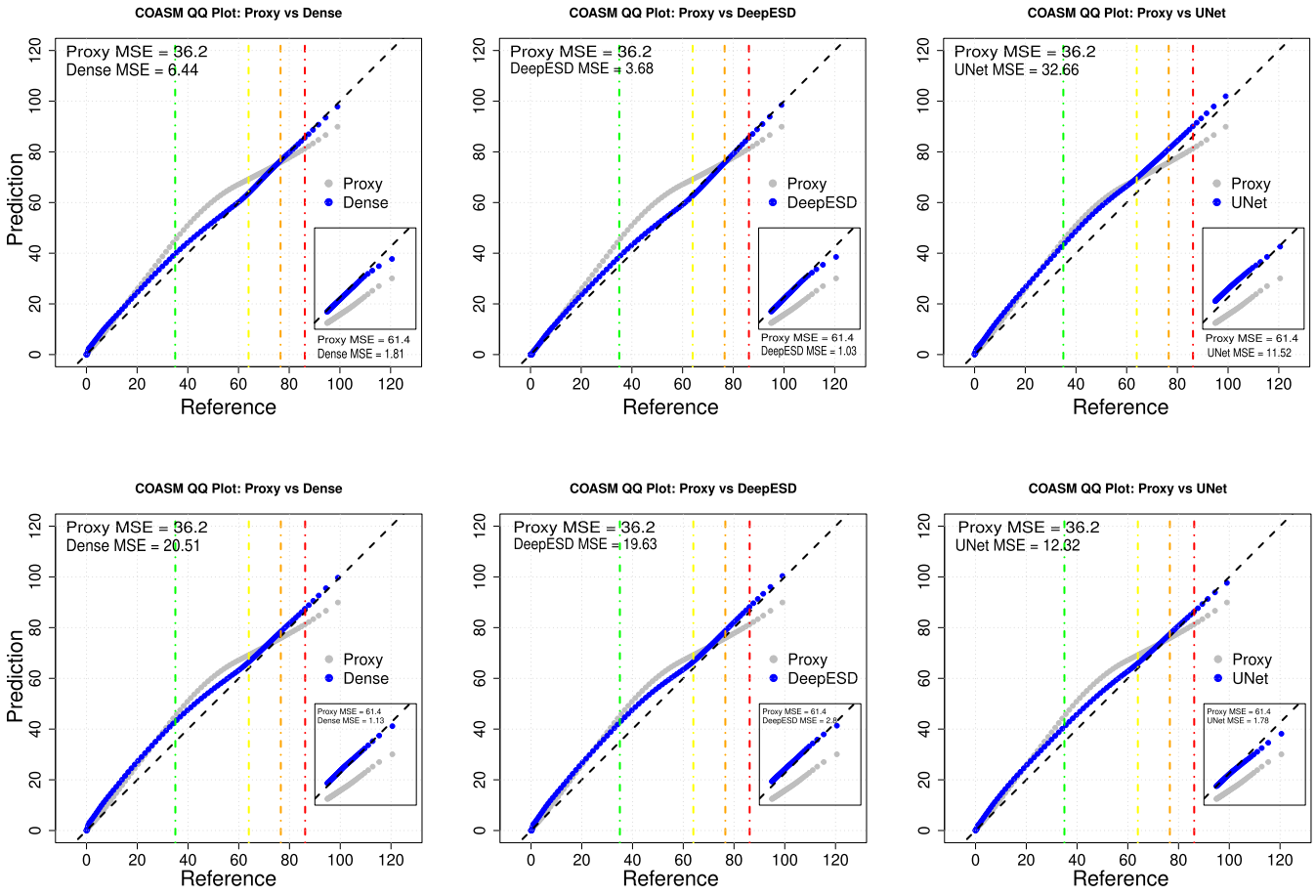

Figure C1. Quantile-Quantile (QQ) plots comparing the Proxy FWI (in grey) and DL model results (in blue). The results for the Coastal Mediterranean (COASM) (Figure A1 for the subregion boundaries). Dash-dotted vertical lines indicate different fire danger levels established by the Agencia Estatal de Meteorología (AEMET). Fire danger levels are defined as follows: *low* (below the green line), *moderate* (between the green and yellow lines), *high* (between the yellow and orange lines), *very high* (between the orange and red lines), and *extreme* (above the red line). The Mean Squared Error (MSE) for both results is displayed in the top right corner. A box in the bottom right provides an amplified view of QQ plot focused on values above the 95th percentile (extreme danger), along with the corresponding MSE values under the box. First row indicates the results for models trained with MSE, and the second provides the models with ASYM as loss function.

Appendix D: Sensitivity analysis of deep learning models and evaluation of other predictors sets

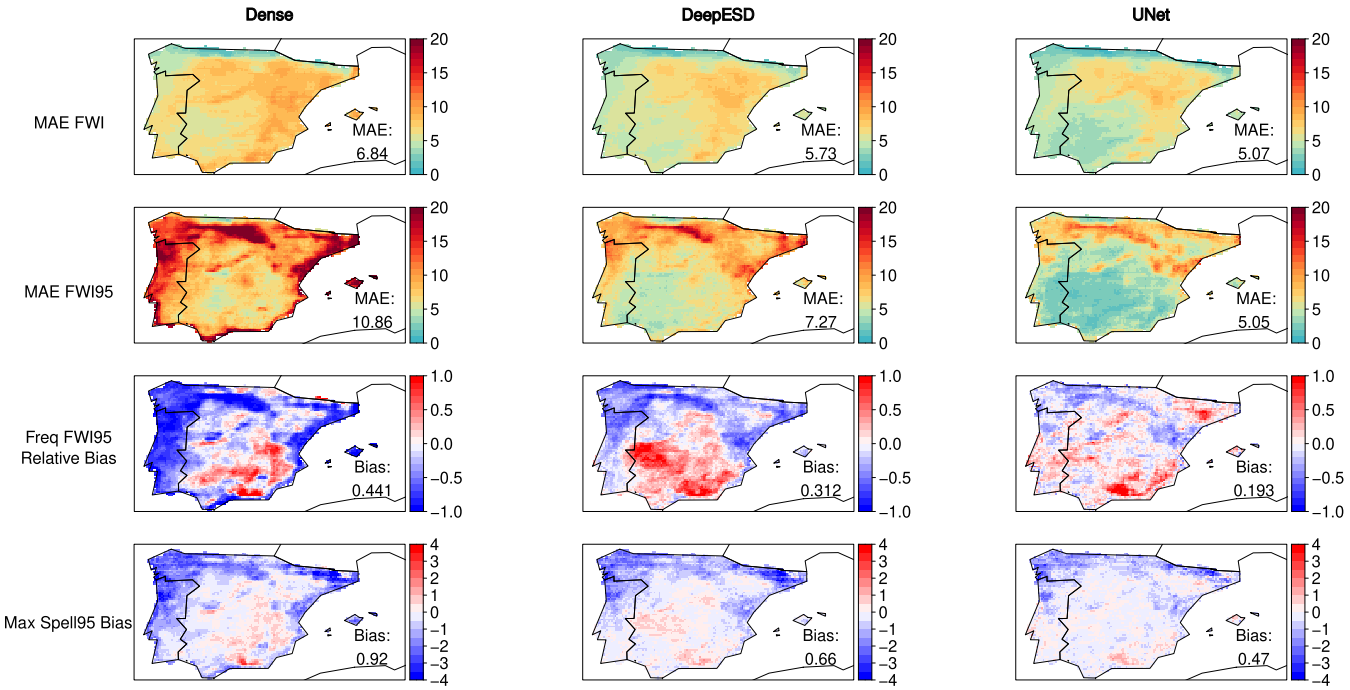

Figure D1. Results from the Dense, DeepESD and U-Net model trained with 12UTC input variables (temperature, 24h-accumulated precipitation, relative humidity and wind speed). The maps display differences relative to the reference FWI for the fire season (June–September) during the test period (2018–2021) for the validation indices. The MAE value inside the map represents the spatially aggregated mean absolute error of the deep learning predictions with respect to the FWI reference, while Bias denotes the spatially averaged bias in absolute value.

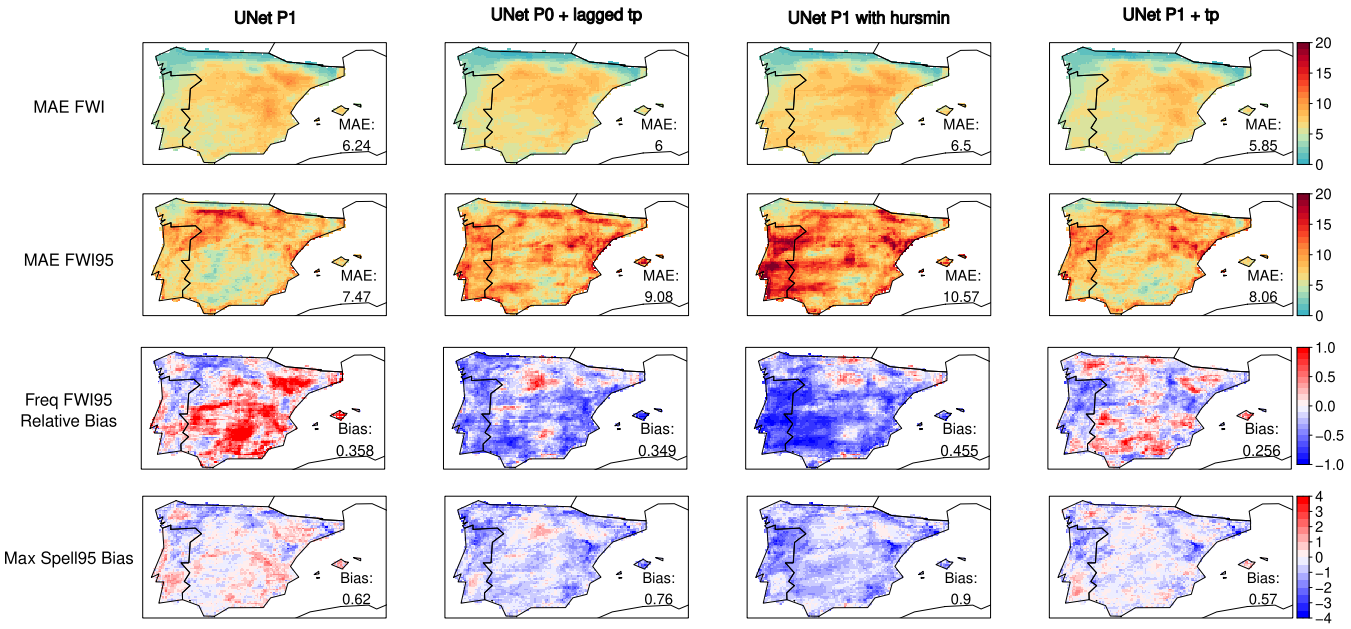

Figure D2. Results from the U-Net model trained with the P1 predictors set (see Table 1), trained with P0 and adding the 24-h lagged precipitation, trained with P1 adding minimum relative humidity instead of daily mean relative humidity and trained with P1 adding precipitation. The maps display differences relative to the reference FWI for the fire season (June–September) during the test period for different validation indices. MAE represents the spatially aggregated mean absolute error of the deep learning predictions with respect to the FWI reference, while Bias denotes the spatially averaged bias in absolute value.

Author contributions. O.M., J.B. and J.B-M. were responsible for the development and implementation of the code used in the experiments. O.M, J.B., J.M.G and J.B-M. contributed to the experimental framework design. All authors contributed to the analysis and interpretation of the results. All authors contributed to manuscript writing and they reviewed and approved the final manuscript.

Competing interests. The authors declare that no competing interests are present.

Acknowledgements. We thank our colleague, Dr. Jose Abad, for his insightful and fruitful discussions on deep learning model configuration and explainability approaches. O.M. has received research support from grant PRE2021-100292 funded by MCIN/AEI /10.13039/501100011033, as part of the R+D+i project CORDyS (PID2020-116595RB-I00) with funding from the Spanish Ministry of Science MCIN/AEI/10.13039/501100011033. J.M.G. and J.B. acknowledge funding by the Ministry for the Ecological Transition and the Demographic Challenge (MITECO) and the European Commission NextGenerationEU (Regulation EU 2020/2094), through CSIC's Interdisciplinary Thematic Platform Clima (PTI-Clima). J.B. has received research support from Grant PID2023-149997OA-

I00 (PROTECT Project) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. P.M.M.S. acknowledges project DHEFEUS (https://doi.org/10.54499/2022.09185.PTDC) and UID/50019/2025 and LA/P/0068/2020 https://doi.org/10.54499/LA/P/0068/2020).

References

Abatzoglou, J. T., Williams, A. P., Boschetti, L., Zubkova, M., and Kolden, C. A.: Global patterns of interannual climate–fire relationships, Global Change Biology, 24, 5164–5175, https://doi.org/10.1111/gcb.14405, 2018.

Baño-Medina, J., Manzanas, R., Cimadevilla, E., Fernández, J., González-Abad, J., Cofiño, A. S., and Gutiérrez, J. M.: Downscaling multi-model climate projection ensembles with deep learning (DeepESD): contribution to CORDEX EUR-44, Geoscientific Model Development, 15, 6747–6758, https://doi.org/10.5194/gmd-15-6747-2022, 2022.

Baño-Medina, J., Iturbide, M., Fernández, J., and Gutiérrez, J. M.: Transferability and Explainability of Deep Learning Emulators for Regional Climate Model Projections: Perspectives for Future Applications, Artificial Intelligence for the Earth Systems, 3, https://doi.org/10.1175/AIES-D-23-0099.1, 2024.

Baño-Medina, J., Sengupta, A., Doyle, J. D., Reynolds, C. A., Watson-Parris, D., and Monache, L. D.: Are AI weather models learning atmospheric physics? A sensitivity analysis of cyclone Xynthia, npj Climate and Atmospheric Science, 8, 1–9, https://doi.org/10.1038/s41612-025-00949-6, 2025.

Bedia, J., Herrera, S., Camia, A., Moreno, J. M., and Gutierrez, J. M.: Forest Fire Danger Projections in the Mediterranean using ENSEMBLES Regional Climate Change Scenarios, Climatic Change, 122, 185–199, https://doi.org/10.1007/s10584-013-1005-z, 2014.

Bedia, J., Herrera, S., Gutierrez, J., Benali, A., Brands, S., Mota, B., and Moreno, J.: Global patterns in the sensitivity of burned area to fire-weather: implications for climate change, Agricultural and Forest Meteorology, 214–215, 369–379, https://doi.org/10.1016/j.agrformet.2015.09.002, 2015.

Bedia, J., Golding, N., Casanueva, A., Iturbide, M., Buontempo, C., and Gutiérrez, J. M.: Seasonal predictions of Fire Weather Index: Paving the way for their operational applicability in Mediterranean Europe, Climate Services, 9, 101–110, https://doi.org/10.1016/j.cliser.2017.04.001, 2018.

Bedia, J., Baño-Medina, J., Legasa, M. N., Iturbide, M., Manzanas, R., Herrera, S., Casanueva, A., San-Martín, D., Cofiño, A. S., and Gutiérrez, J. M.: Statistical downscaling with the downscaleR package (v3.1.0): contribution to the VALUE intercomparison experiment, Geoscientific Model Development, 13, 1711–1735, https://doi.org/10.5194/gmd-13-1711-2020, 2020.

Bento, V. A., Lima, D. C. A., Santos, L. C., Lima, M. M., Russo, A., Nunes, S. A., DaCamara, C. C., Trigo, R. M., and Soares, P. M. M.: The future of extreme meteorological fire danger under climate change scenarios for Iberia, Weather and Climate Extremes, 42, 100 623, https://doi.org/10.1016/j.wace.2023.100623, 2023.

Bowman, D. M. J. S., Balch, J. K., Artaxo, P., Bond, W. J., Carlson, J. M., Cochrane, M. A., D'Antonio, C. M., DeFries, R. S., Doyle, J. C., Harrison, S. P., Johnston, F. H., Keeley, J. E., Krawchuk, M. A., Kull, C. A., Marston, J. B., Moritz, M. A., Prentice, I. C., Roos, C. I., Scott, A. C., Swetnam, T. W., van der Werf, G. R., and Pyne, S. J.: Fire in the Earth System, Science, 324, 481–484, https://doi.org/10.1126/science.1163886, 2009.

Bowman, D. M. J. S., Williamson, G. J., Abatzoglou, J. T., Kolden, C. A., Cochrane, M. A., and Smith, A. M. S.: Human exposure and sensitivity to globally extreme wildfire events, Nature Ecology & Evolution, 1, 0058, https://doi.org/10.1038/s41559-016-0058, 2017.

Buontempo, C., Burgess, S. N., Dee, D., Pinty, B., Thépaut, J.-N., Rixen, M., Almond, S., Armstrong, D., Brookshaw, A., Alos, A. L., Bell, B., Bergeron, C., Cagnazzo, C., Comyn-Platt, E., Damasio-Da-Costa, E., Guillory, A., Hersbach, H., Horányi, A., Nicolas, J., Obregon, A., Ramos, E. P., Raoult, B., Muñoz-Sabater, J., Simmons, A., Soci, C., Suttie, M., Vamborg, F., Varndell, J., Vermoote, S., Yang, X., and Marcilla, J. G. d.: The Copernicus Climate Change Service: Climate Science in Action, Bulletin of the American Meteorological Society, 103, E2669–E2687, https://doi.org/10.1175/BAMS-D-21-0315.1, 2022.

Bushenkova, A., Soares, P. M. M., Johannsen, F., and Lima, D. C. A.: Towards an improved representation of the urban heat island effect : A multi-scale application of XGBoost for madrid, Urban Climate, 55, 101 982, https://doi.org/10.1016/j.uclim.2024.101982, 2024.

Camia, A. and Amatulli, G.: Weather Factors and Fire Danger in the Mediterranean, in: Earth Observation of Wildland Fires in Mediterranean Ecosystems, edited by Chuvieco, E., pp. 71–82, Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-01754-4_6, 2009.

Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., Ananthakrishnan, R., Miller, N., Denvil, S., Morgan, M., Pobre, Z., Bell, G. M., Doutriaux, C., Drach, R., Williams, D., Kershaw, P., Pascoe, S., Gonzalez, E., Fiore, S., and Schweitzer, R.: The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data, Future Generation Computer Systems, 36, 400–417, https://doi.org/10.1016/j.future.2013.07.002, 2014.

Costa, H., De Rigo, D., Libertà, G., Houston Durrant, T., and San-Miguel-Ayanz, J.: European wildfire danger and vulnerability in a changing climate – Towards integrating risk dimensions – JRC PESETA IV project – Task 9 - forest fires, European Commission and Joint Research Centre, Publications Office of the European Union, https://doi.org/doi/10.2760/46951, 2020.

DaCamara, C. C., Bento, V. A., Nunes, S. A., Lemos, G., Soares, P. M. M., and Trigo, R. M.: Impacts of fire prevention strategies in a changing climate: an assessment for Portugal, Environmental Research: Climate, 3, 045 002, https://doi.org/10.1088/2752-5295/ad574f, publisher: IOP Publishing, 2024.

Di Giuseppe, F., Vitolo, C., Krzeminski, B., Barnard, C., Maciel, P., and San-Miguel, J.: Fire Weather Index: the skill provided by the European Centre for Medium-Range Weather Forecasts ensemble prediction system, Natural Hazards and Earth System Sciences, 20, 2365–2378, https://doi.org/10.5194/nhess-20-2365-2020, 2020.

Doury, A., Somot, S., and Gadat, S.: On the suitability of a convolutional neural network based RCM-emulator for fine spatio-temporal precipitation, Climate Dynamics, 62, 8587–8613, https://doi.org/10.1007/s00382-024-07350-8, 2024.

Dramsch, J. S., Kuglitsch, M. M., Fernández-Torres, M. A., Toreti, A., Albayrak, R. A., Nava, L., Ghaffarian, S., Cheng, X., Ma, J., Samek, W., Venguswamy, R., Koul, A., Muthuregunathan, R., and Hrast Essenfelder, A.: Explainability can foster trust in artificial intelligence in geoscience, Nature Geoscience, 18, 112–114, https://doi.org/10.1038/s41561-025-01639-x, 2025.

Duffy, K., Vandal, T. J., Wang, W., Nemani, R. R., and Ganguly, A. R.: A Framework for Deep Learning Emulation of Numerical Models With a Case Study in Satellite Remote Sensing, IEEE Transactions on Neural Networks and Learning Systems, 34, 3345–3356, https://doi.org/10.1109/TNNLS.2022.3169958, conference Name: IEEE Transactions on Neural Networks and Learning Systems, 2023.

Dupuy, J.-l., Fargeon, H., Martin-StPaul, N., Pimont, F., Ruffault, J., Guijarro, M., Hernando, C., Madrigal, J., and Fernandes, P.: Climate change impact on future wildfire danger and activity in southern Europe: a review, Annals of Forest Science, 77, 1–24, https://doi.org/10.1007/s13595-020-00933-5, 2020.

Fugioka, F. M., Gill, A., Viegas, D. X., and Wotton, B.: Fire Danger and Fire Behavior Modeling Systems in Australia, Europe, and North America, in: Developments in Environmental Science, edited by Bytnerowicz, A., Arbaugh, M., Riebau, A., and Andersen, C., Elsevier B.V., The Netherlands, 2009.

Gincheva, A., Pausas, J. G., Torres-Vázquez, M. A., Bedia, J., Vicente-Serrano, S. M., Abatzoglou, J. T., Sánchez-Espigares, J. A., Chuvieco, E., Jerez, S., Provenzale, A., Trigo, R. M., and Turco, M.: The Interannual Variability of Global Burned Area Is Mostly Explained by Climatic Drivers, Earth's Future, 12, e2023EF004 334, https://doi.org/10.1029/2023EF004334, 2024.

Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256, JMLR Workshop and Conference Proceedings, https://proceedings.mlr.press/v9/glorot10a.html, iSSN: 1938-7228, 2010.

González-Abad, J. and Gutiérrez, J. M.: Are Deep Learning Methods Suitable for Downscaling Global Climate Projections? Review and Intercomparison of Existing Models, https://doi.org/10.48550/arXiv.2411.05850, arXiv:2411.05850 [physics], 2024.

González-Abad, J., Baño-Medina, J., and Gutiérrez, J. M.: Using Explainability to Inform Statistical Downscaling Based on Deep Learning Beyond Standard Validation Approaches, Journal of Advances in Modeling Earth Systems, 15, e2023MS003 641, https://doi.org/10.1029/2023MS003641, 2023.

Herrera, S., Bedia, J., Gutierrez, J. M., Fernandez, J., and Moreno, J. M.: On the projection of future fire danger conditions with various instantaneous/mean-daily data sources, Climatic Change, 118, 827–840, https://doi.org/10.1007/s10584-012-0667-2, 2013.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hogan, R. J. and Mason, I. B.: Deterministic Forecasts of Binary Events, chap. 3, pp. 31–59, John Wiley & Sons, Ltd, https://doi.org/10.1002/9781119960003.ch3, 2011.

Iturbide, M., Bedia, J., Herrera, S., Baño-Medina, J., Fernández, J., Frías, M., Manzanas, R., San-Martín, D., Cimadevilla, E., Cofiño, A., and Gutiérrez, J.: The R-based climate4R open framework for reproducible climate data access and post-processing, Environmental Modelling & Software, 111, 42–54, https://doi.org/10.1016/j.envsoft.2018.09.009, 2019.

Jain, P., Castellanos-Acuna, D., Coogan, S. C. P., Abatzoglou, J. T., and Flannigan, M. D.: Observed increases in extreme fire weather driven by atmospheric humidity and temperature, Nature Climate Change, 12, 63–70, https://doi.org/10.1038/s41558-021-01224-1, publisher: Nature Publishing Group, 2022.

Johannsen, F., Soares, P. M. M., and Langendijk, G. S.: On the deep learning approach for improving the representation of urban climate: The Paris urban heat island and temperature extremes, Urban Climate, 56, 102 039, https://doi.org/10.1016/j.uclim.2024.102039, 2024.

Jolly, W. M., Cochrane, M. A., Freeborn, P. H., Holden, Z. A., Brown, T. J., Williamson, G. J., and Bowman, D. M. J. S.: Climate-induced variations in global wildfire danger from 1979 to 2013, Nature Communications, 6, 7537, https://doi.org/10.1038/ncomms8537, 2015.

Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., Fernández-Torres, M.-A., and Carvalhais, N.: Wildfire Danger Prediction and Understanding With Deep Learning, Geophysical Research Letters, 49, e2022GL099 368, https://doi.org/10.1029/2022GL099368, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2022GL099368, 2022.

Lawson, B. D. and Armitage, O. B.: Weather Guide for the Canadian Forest Fire Danger Rating System, Tech. rep., Canadian Forest Service, 2008.

Li, S., Wan, H., Yu, Q., and Wang, X.: Downscaling of ERA5 reanalysis land surface temperature based on attention mechanism and Google Earth Engine, Scientific Reports, 15, 675, https://doi.org/10.1038/s41598-024-83944-w, 2025.

Mamalakis, A., Barnes, E. A., and Hurrell, J. W.: Using Explainable Artificial Intelligence to Quantify "Climate Distinguishability" After Stratospheric Aerosol Injection, Geophysical Research Letters, 50, e2023GL106 137, https://doi.org/https://doi.org/10.1029/2023GL106137, 2023.

Matteo, A., Garnés-Morales, G., Moreno, A., Andreia, R., Azorín-Molina, C., Bedia, J., Giuseppe, F. D., Dunn, R. J. H., Herrera, S., Provenzale, A., Quilcaille, Y., Vázquez, M. A. T., and Turco, M.: Challenges in assessing Fire Weather changes in a warming climate, https://doi.org/10.48550/arXiv.2503.01818, arXiv:2503.01818 [physics], 2025.

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T.: Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning, Bulletin of the American Meteorological Society, 100, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 2019.

Mirones, O., Baño-Medina, J., Brands, S., and Bedia, J.: Toward Spatio-Temporally Consistent Multi-Site Fire Danger Downscaling With Explainable Deep Learning, Journal of Geophysical Research: Machine Learning and Computation, 2, e2024JH000331, https://doi.org/10.1029/2024JH000331, 2025.

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, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth System Science Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Pausas, J. G. and Keeley, J. E.: Wildfires and global change, Frontiers in Ecology and the Environment, 19, 387–395, https://doi.org/10.1002/fee.2359, 2021.

Prince, S. J.: Understanding Deep Learning, The MIT Press, http://udlbook.com, 2023.

Rampal, N., Hobeichi, S., Gibson, P. B., Abramowitz, G., Beucler, T., Gonz, J., Chapman, W., Harder, P., and Guti, M.: Enhancing Regional Climate Downscaling through Advances in Machine Learning, IEEE Transactions on Neural Networks and Learning Systems, 3, 2024.

San-Miguel-Ayanz, J., Schulte, E., Schmuck, G., and Camia, A.: The European Forest Fire Information System in the context of environmental policies of the European Union, Forest Policy and Economics, 29, 19–25, https://doi.org/10.1016/j.forpol.2011.08.012, 2013.

Santos, L. C., Lima, M. M., Bento, V. A., Nunes, S. A., DaCamara, C. C., Russo, A., Soares, P. M. M., and Trigo, R. M.: An Evaluation of the Atmospheric Instability Effect on Wildfire Danger Using ERA5 over the Iberian Peninsula, Fire, 6, 120, https://doi.org/10.3390/fire6030120, number: 3 Publisher: Multidisciplinary Digital Publishing Institute, 2023.

Soares, P. M. M., Johannsen, F., Lima, D. C. A., Lemos, G., Bento, V. A., and Bushenkova, A.: High-resolution downscaling of CMIP6 Earth system and global climate models using deep learning for Iberia, Geoscientific Model Development, 17, 229–259, https://doi.org/10.5194/gmd-17-229-2024, publisher: Copernicus GmbH, 2024.

Stocks, B. J., Lawson, B. D., Alexander, M. E., Wagner, C. E. V., McAlpine, R. S., Lynham, T. J., and Dubé, D. E.: The Canadian Forest Fire Danger Rating System: An Overview, The Forestry Chronicle, 65, 450–457, https://doi.org/10.5558/tfc65450-6, publisher: Canadian Institute of Forestry, 1989.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic attribution for deep networks, in: International conference on machine learning, pp. 3319–3328, PMLR, 2017.

Toms, B. A., Kashinath, K., Prabhat, and Yang, D.: Testing the reliability of interpretable neural networks in geoscience using the Madden–Julian oscillation, Geoscientific Model Development, 14, 4495–4508, https://doi.org/10.5194/gmd-14-4495-2021, 2021.

van Wagner, C. E.: Development and structure of the Canadian Forest Fire Weather Index System, Tech. rep., Minister of Supply and Services Canada, Ottawa, https://ostrnrcan-dostrncan.canada.ca/handle/1845/228434, 1987.

Wotton, B. M.: Interpreting and using outputs from the Canadian Forest Fire Danger Rating System in research applications, Environmental and Ecological Statistics, 16, 107–131, https://doi.org/10.1007/s10651-007-0084-2, 2009.

Yang, R., Hu, J., Li, Z., Mu, J., Yu, T., Xia, J., Li, X., Dasgupta, A., and Xiong, H.: Interpretable machine learning for weather and climate prediction: A review, Atmospheric Environment, 338, 120797, https://doi.org/10.1016/j.atmosenv.2024.120797, 2024.

Yu, G., Feng, Y., Wang, J., and Wright, D. B.: Performance of Fire Danger Indices and Their Utility in Predicting Future Wildfire Danger Over the Conterminous United States, Earth's Future, 11, e2023EF003 823, https://doi.org/10.1029/2023EF003823, 2023.