We sincerely thank the reviewer for taking the time to review our manuscript. We believe their feedback has improved the clarity of the manuscript and overall quality of this work.

In this study, the authors demonstrate the ability of deep learning (DL) models to emulate the Fire Weather Index (FWI) at 12 UTC. Specifically, three DL models are trained using either daily means or proxy data (as in Bedia et al., 2014) of weather variables relevant to FWI computation, to produce noon-time FWI estimates based on ERA5-Land. The authors also apply interpretability techniques to rank input variables according to their relevance in producing the FWI output. They find that, in high and extreme FWI scenarios, 24-hour accumulated precipitation is not needed to obtain accurate FWI values.

I appreciate the motivation behind this work and the authors' methodology. The results are compelling and well presented. However, I believe a deeper analysis in certain areas would significantly enhance the value of the paper.

<u>Major Comment 1: Generalization to datasets other than ERA5-Land</u>

One of the main motivations of this work is to emulate reference FWI conditions (i.e., those computed using 12 UTC weather variables and 24-hour accumulated precipitation) using proxy data from datasets that typically provide only daily information, such as climate model outputs. However, the authors do not show an example of applying their DL models to such external datasets.

Given that DL models are often sensitive to the data distribution used during training, applying the trained models to daily means from a dataset different from ERA5-Land (e.g., GCMs or other reanalysis) may yield inaccurate FWI emulations. Potential issues include discrepancies in statistical properties (e.g., mean, variance, extremes), spatial resolution (important for CNNs), or temporal characteristics.

I suggest the authors assess how their models perform when applied to an alternative dataset to emphasize the potential of this approach for correcting FWI estimates in climate simulations lacking noon-time fields.

We appreciate the referee's thoughtful comment and fully agree that assessing the performance of the proposed models on other datasets, such as GCM outputs, is an important future step. In this study, however, we focus solely on emulation: that is, learning the transfer function with a DL model using the same database for both predictor set and predictand (ERA5-Land). Our primary goals are to evaluate the ability of DL models to emulate a multivariate index (FWI, in this case) from linked variables, setting a validation framework and looking for a reliable set up for the DL architectures and illustrating an intercomparison between the architectures, rather than to assess transferability across datasets (e.g., to other reanalyses or GCMs), which lies beyond the scope of this paper.

Nevertheless, we recognize the importance of transferability and are actively preparing a follow-up manuscript in which the proposed models will be applied to GCM outputs. In that

work, we will conduct a more detailed analysis of model performance on datasets with different statistical and temporal characteristics than ERA5-Land, and discuss the implications for correcting FWI estimates in climate simulations that lack noon-time meteorological fields. As noted earlier, we are also working on extending this line of research to GCM-based downscaling studies.

It is important to note that applying these models to other datasets, such as GCMs, requires careful consideration of statistical differences (e.g., biases in mean, variance, extremes, or spatial resolution), which would necessitate additional preprocessing (such as bias correction) and extensive validation. A rigorous evaluation of transferability therefore requires a more comprehensive and systematic study than can reasonably be accommodated within the scope of this manuscript.

To summarize, our study should be regarded as a case study on the ability of DL models to emulate FWI. While we see strong potential for their application beyond ERA5-Land, we consider it more appropriate to address transferability to other datasets (e.g., GCMs or alternative reanalyses) in future work. Given the already substantial length of the manuscript and the aforementioned scope, incorporating such an analysis here would not be feasible.

<u>Major Comment 2: Temporal evaluation of FWI estimations</u>

The authors conclude that their DL models capture both spatial and temporal variability of the reference FWI better than traditional proxy methods, and improve the detection of high-risk events. While the spatial evaluation is clearly presented, the paper does not seem to explicitly evaluate the temporal aspects of the DL-predicted FWI.

Beyond the Max Spell analysis, I suggest comparing the seasonal cycle of the reference FWI, proxy FWI, and DL-predicted FWI. This would help assess whether the models maintain consistent accuracy across different parts of the year or under seasonal biases. I also recommend extending the test dataset beyond the current 3-year window—e.g., from 2018 to 2024—to ensure more robust temporal assessment.

We thank the reviewer for the suggestion regarding the evaluation of temporal aspects of the FWI predictions. 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, U-Net) per climatological region, as shown in Figure 1. Each boxplot represents the distribution of FWI values for a given month over the 2012–2021 period, capturing both the median and the spread of the values.

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 the different regions. The median and interquartile ranges of the DL predictions closely follow the reference FWI throughout the seasonal cycle, including the high fire danger summer months,

whereas the traditional proxy FWI tends to slightly underestimate extremes during peak months (June–September). This analysis confirms that the DL models maintain consistent accuracy across different parts of the year and do not introduce significant seasonal biases.

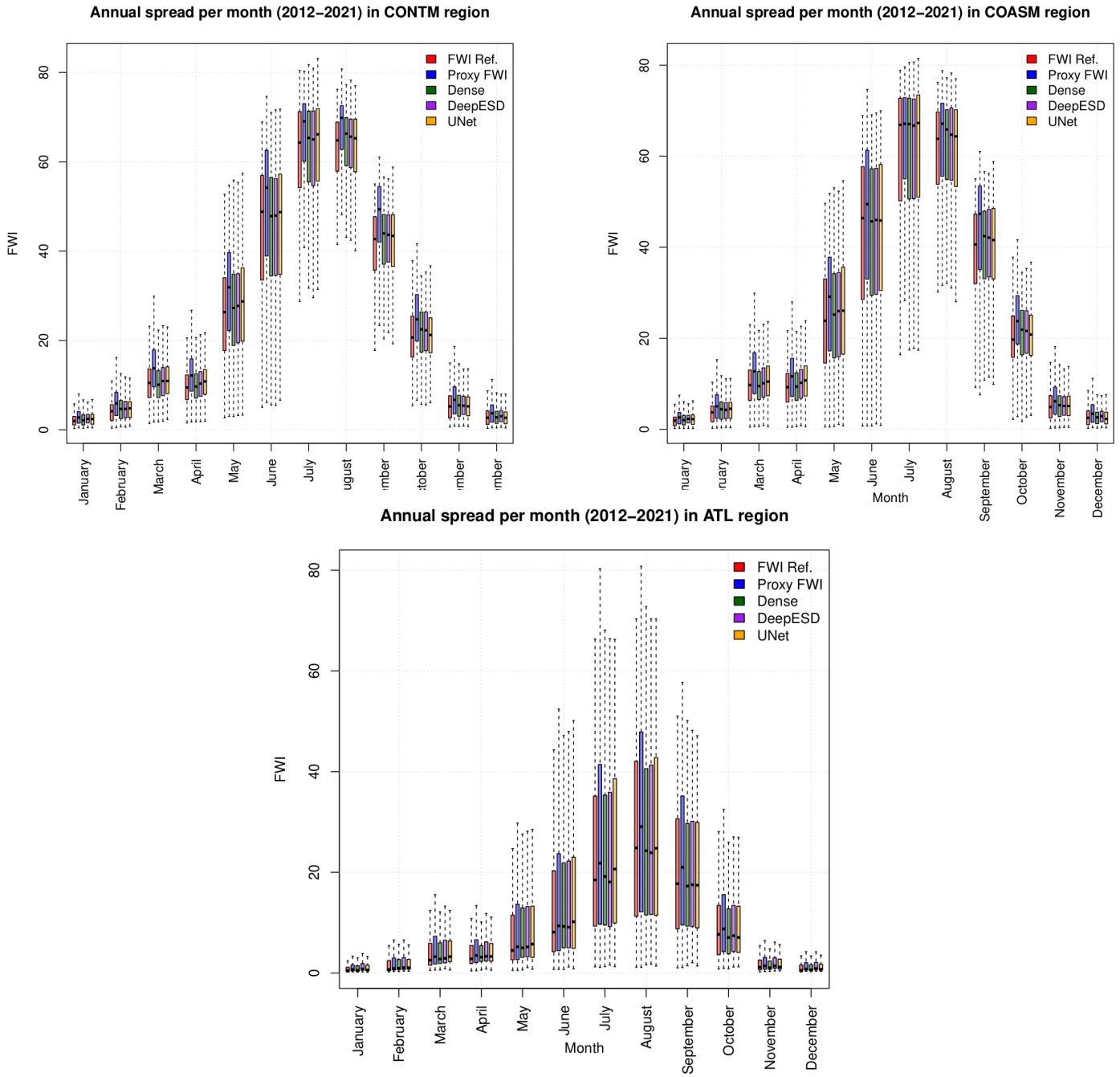

Figure 1: Monthly boxplots for the reference FWI, proxy FWI, and the DL-predicted FWI (Dense, DeepESD, UNet) for CONTM (top-left), COASM (top-right) and ATL (bottom) regions (see Figure A1 of the manuscript to understand the climatological regions). Each boxplot represents the distribution of FWI values for a given month over the test period, capturing both the median and the spread of the values.

The discussion provided in this comment about the temporal evaluation and the Figure 1 have been added to the manuscript in Section 3.2.1.

Moreover, as suggested by the referee, we have extended the test period from 2018–2021 to 2012–2021 in order to provide a more comprehensive evaluation of the method's temporal robustness. By increasing the length of the test period, we are able to assess model performance across a wider range of interannual variability and climatic conditions, which allows for a more rigorous validation. The results obtained for the extended period are consistent with those observed in the original 2018–2021 test window, indicating that the method maintains its accuracy and reliability over a longer timeframe. This outcome reinforces the robustness of our approach and provides additional confidence in the generalizability of the model for FWI emulation.

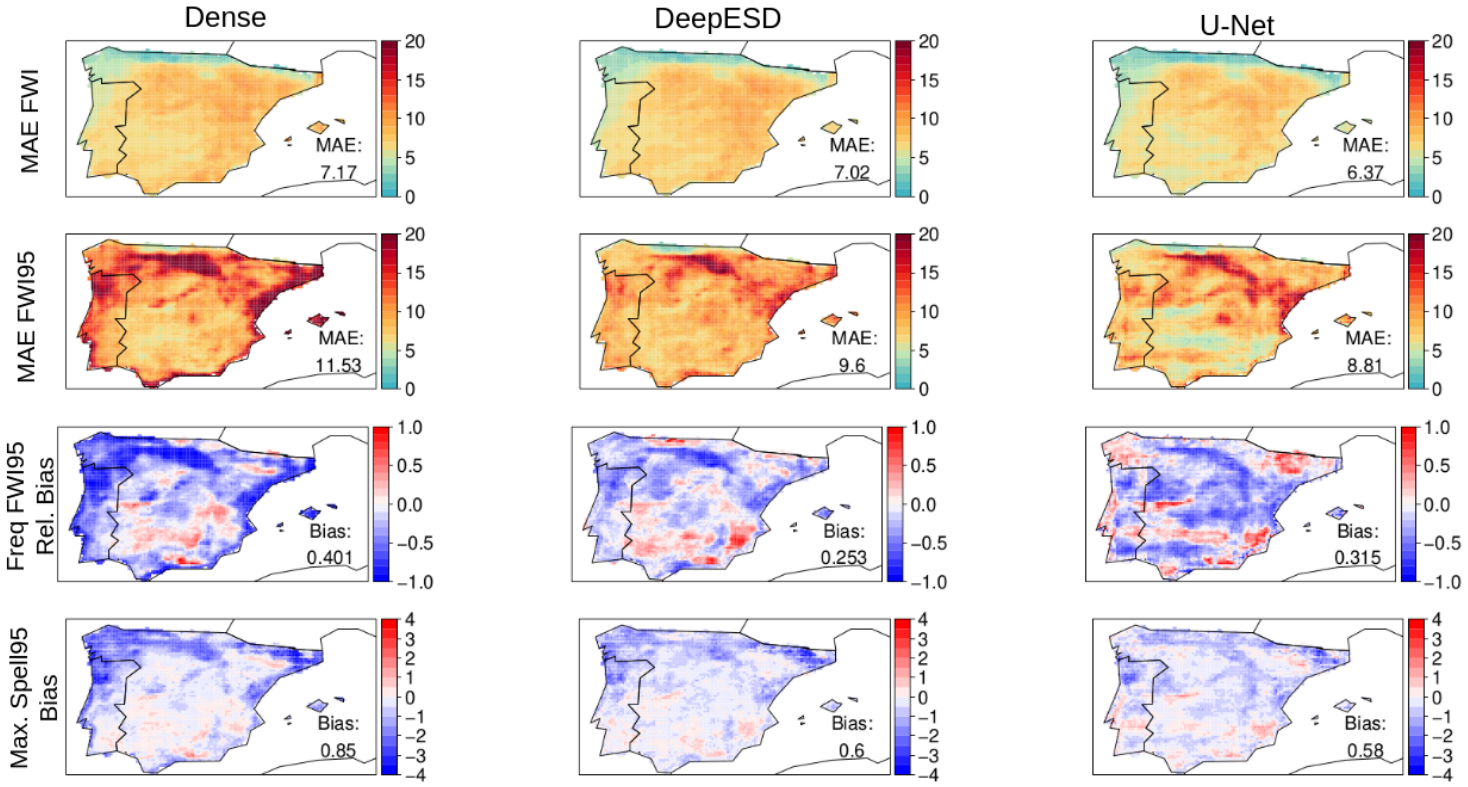

Figure 2: Results from the Dense, DeepESD and U-Net model trained with the P0 predictors set (see Table 1 from manuscript) . The maps display differences relative to the reference FWI for the fire season (June–September) during the test period (2012–2021) for the FWI MAE (first row), FWI95 MAE (second row), Frequency FWI95 Relative Bias (third row) and Maximum annual Spell FWI 95 bias. 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.

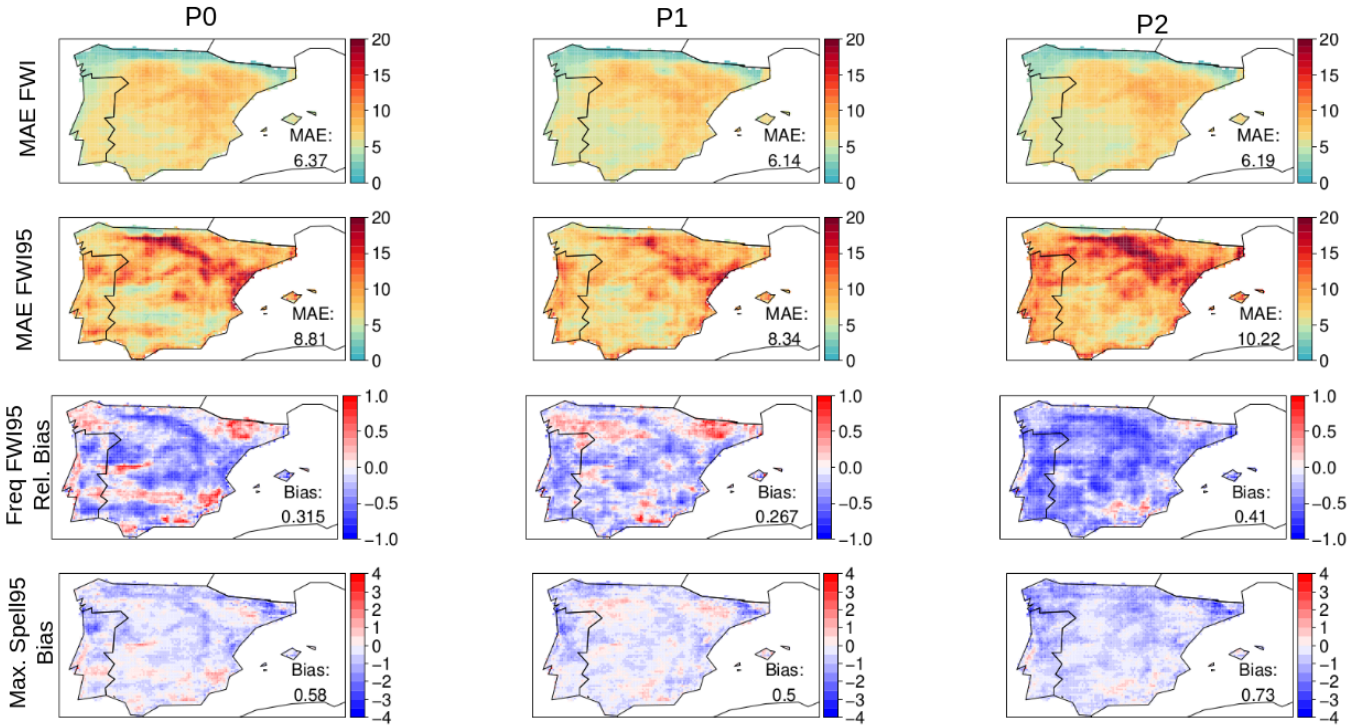

Figure 3: Results from the U-Net model trained with the P0, P1 and predictors sets (see Table 1 from manuscript) . The maps display differences relative to the reference FWI for the fire season (June–September) during the test period (2012–2021) for the FWI MAE (first row), FWI95 MAE (second row), Frequency FWI95 Relative Bias (third row) and Maximum annual Spell FWI 95 bias. 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.

Therefore, as previously commented, we obtain robust results independently of modifications in train (from 1979-2017 to 1979-2011) or test (from 2018-2021 to 2012-2021) period. The values of spatially aggregated MAE and bias of Figure 2 and 3 are consistent with the ones in the manuscript (Figures 3 and 6 respectively). Also the spatial patterns observed in the validation indices maps are similar, with a few exceptions such as the case of the Relative Bias of the FWI95 Frequency for the U-Net model trained with the P0 and P1 pattern. Accordingly, we mention in the new version of the manuscript in Section 2.5 that "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".

<u>Minor 1:</u>

The authors justify using the term "reference FWI" instead of "ground truth" because ERA5-Land is not observational. Since the DL models are trained on ERA5-Land, they likely inherit its biases. I suggest briefly discussing known limitations of ERA5-Land compared to observations, especially if no comparison with observed FWI is

included. This is specially important considering major comment 1 in which the ERA5Land's biases learned by the model might be propagate to other models.

Thank you for the comment. We will include in the discussion the limitations of ERA5-Land compared to the observations adding as appendix the biases between observational FWI data from the Spanish Agency of Meteorology (AEMET) and the FWI resulting from ERA5-Land computations. Here, we present the assessment about the limitations of ERA5-Land FWI compared with the observational data in terms of FWI and FWI95 climatologies biases.

In the new Appendix B: ERA5-Land limitations we have added the following text and Figure:

"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. 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)."

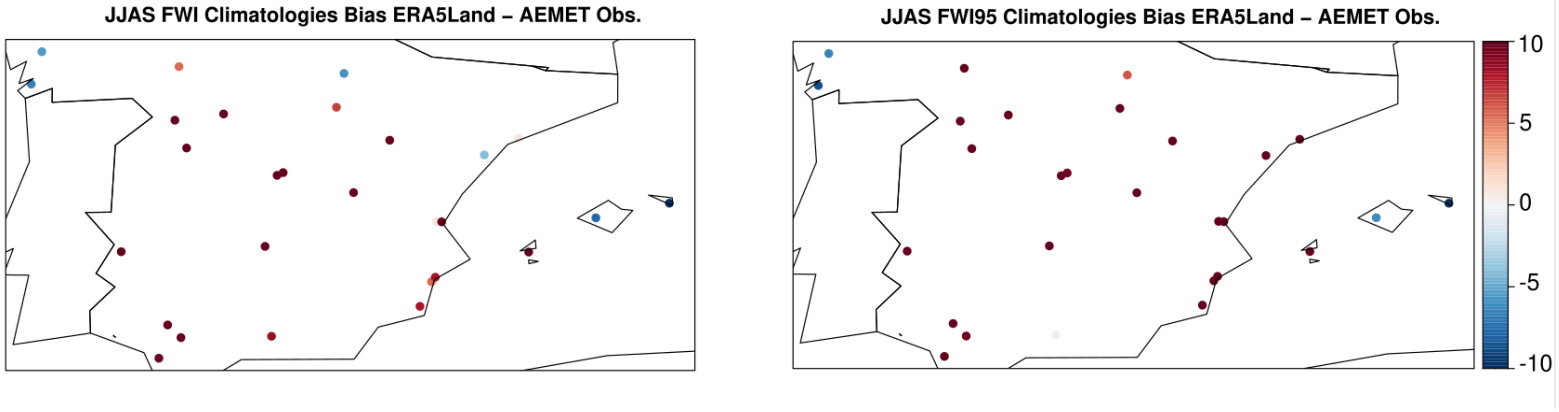

Figure 4: Biases between observational FWI data from the Spanish Agency of Meteorology (AEMET) and the FWI resulting from ERA5-Land computations. The map in the left indicates bias for the mean FWI, while the map in the right indicates it for the FWI95.

<u>Minor 2:</u>

Please provide the actual thresholds used by the Spanish Meteorological Agency (AEMET) for fire danger classification, as these can vary by country and are important for interpretation.

Thank you for your comment. We will include the AEMET fire danger threshold in the manuscript, as it is necessary for understanding the FWI magnitudes associated with each category. We are also considering adding a table to the manuscript that briefly summarizes the thresholds for each category. We have added this table in the manuscript in Section 2.5.

Table: 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

Minor 3:

I suggest moving Figure 1 to the Supplementary Information, as similar architectures have already been described in previous literature.

We thank the referee for the suggestion. Although it is true that similar architectures have been described in the previous literature, we strongly believe that including this figure makes the study more self-contained and improves reader comprehension. For readers who are not familiar with Deep Learning, it may be difficult to fully grasp the architecture from the textual description alone. Moreover, having the figure available directly in the manuscript makes the presentation clearer and more accessible.

Minor 4:

Consider merging Figures 2 and 3 to highlight the comparison between reference FWI, proxy FWI, and the DL emulators in a more compact and interpretable format.

We appreciate the referee's observation. However, we believe that merging Figures 2 and 3 would result in a very dense and large figure, reducing clarity. Additionally, we think the figures should remain separate because, first, we present our reference—the proxy commonly used in the literature—and then the bias between these two approaches. Furthermore, throughout the manuscript, validation figures such as Figures 3 and 6 are used to illustrate model bias or errors with respect to the true target values. In our view, this separation makes it easier to follow the narrative, identify issues, and clearly understand the results.

Minor 5:

It's unclear why the Freq. FWI95 for 12 UTC is not shown in Figure 1, even though it is later used to compute biases. Including it would help clarify the comparison.

Thank you for your comment. Although the frequency of FWI95 events is a key metric in validation, it is not explicitly shown in Figure 2 because, by construction, each grid point in the reference data records exactly 5% of days above its local 95th percentile threshold during the season. In other words, the value in each grid across the spatial map is 0.05. We have clarified this in the new version of the manuscript:

Section 3.1: "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."

Minor 6:

Could the overestimation of Freq. FWI95 by the UNet be explained by a general overestimation of FWI in this model, as suggested by the scatter plot? Are all three DL models trained on exactly the same days and years?

Yes, all three DL models (Dense, DeepESD, and UNet) were trained on exactly the same days and years and also with the same parameters setup, ensuring a fair comparison. Regarding the overestimation of the frequency of FWI95 by the UNet, this effect is due to an error in the calculation of the Figure. The correct version is the following attached below, therefore now showing a similar spatial pattern than the other DL models:

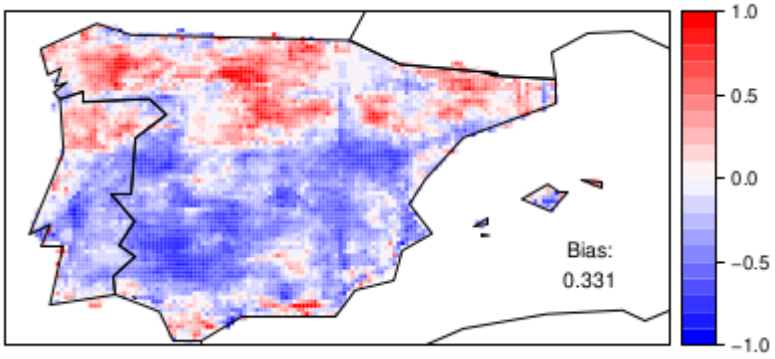

Figure 5: Results from the U-Net model trained with the P0 for the FWI95 Frequency Relative Bias. The maps display differences relative to the reference FWI for the fire season (June–September) during the test period (2018–2021) for the FWI MAE (first row). The Bias shown inside the panel denotes the spatially averaged bias in absolute value.

We thank the referee for the observation and this problem will be solved modifying the affected figures in the main manuscript. Despite this, the manuscript's storyline remains unchanged

Minor 7:

Please use either "proxy FWI" or "Proxy FWI" consistently throughout the text.

Thank you for your observation. We have replaced "proxy FWI" by "Proxy FWI" to be consistent throughout the text.

Minor 8:

Have you normalized the input data before training the DL models? If so, please specify the normalization method used.

Yes, we normalize our input data following a standardization. We set up the standardization as follows:

$$x_i' = (x_i - \mu_i)/\sigma_i \quad \text{where i is the corresponding gridpoint}$$

, where x' represents the standardized value, x represents the raw (unstandardized) value, μ_i and σ_i represents the mean and standard deviation at gridpoint i. Parameters mu and sigma have been computed relative to the training period 1979-2011 (per gridpoint).

When the model is applied to unseen (test) data, the standardization is performed using the same μ_i and σ_i values derived from the training data. This ensures that the test data is normalized in a way that is consistent with the training data, preventing information leakage and maintaining comparability between training and prediction phases.

A brief explanation has been added to the manuscript in Section 2.4.

Minor 9:

Since your DL architectures do not incorporate temporal dependencies, they may miss the effect of temporal accumulation in the Duff Moisture and Drought codes (e.g., DC, DMC). Why did you choose non-recurrent architectures over those incorporating temporal structure (e.g., LSTMs)?

We thank the referee for the constructive comments. In response, we implemented a ConvLSTM model with a temporal window of seven days (i.e., to predict day i, the model incorporates information from days i–6 through i). Given the substantial computational demands of ConvLSTM training, this implementation was carried out on GPUs. The ConvLSTM model adopts the general U-Net framework described in the manuscript, but substitutes the 2D convolutional layers with 2D ConvLSTM layers. Replicating the full depth of the original U-Net architecture was not feasible due to memory constraints; therefore, the depth of the ConvLSTM model was reduced accordingly.

Figure 6 presents a comparison between the ConvLSTM results and those of the model described in the manuscript for the JJAS season during the test period (2012–2022). The training parameters and conditions were kept identical to ensure comparability. Across the principal validation metrics reported in the paper, the ConvLSTM exhibits inferior performance relative to the other models. This outcome does not necessarily imply that ConvLSTMs are unsuitable for emulating the FWI; rather, it suggests that achieving competitive performance would require further optimization and the design of a more complex ConvLSTM architecture.

Such an investigation is currently underway, as part of ongoing work aimed at optimizing time-dependent models for FWI downscaling applications. However, these efforts fall beyond the scope of the present study, which is focused on FWI emulation. Moreover, we consider that the use of ConvLSTM is not strictly necessary to capture temporal dependencies, as these are inherently embedded in the computation of the FWI itself. As the referee rightly

noted, several components of the FWI—such as the Drought Code (DC) and the Duff Moisture Code (DMC)—explicitly account for the temporal evolution of moisture and drought conditions. Therefore, the FWI, as the target variable, already encapsulates this temporal characterization. Consequently, we consider that the absence of explicit temporal modeling in the network architecture is not critical, as the models are learning a function that implicitly includes these dependencies.

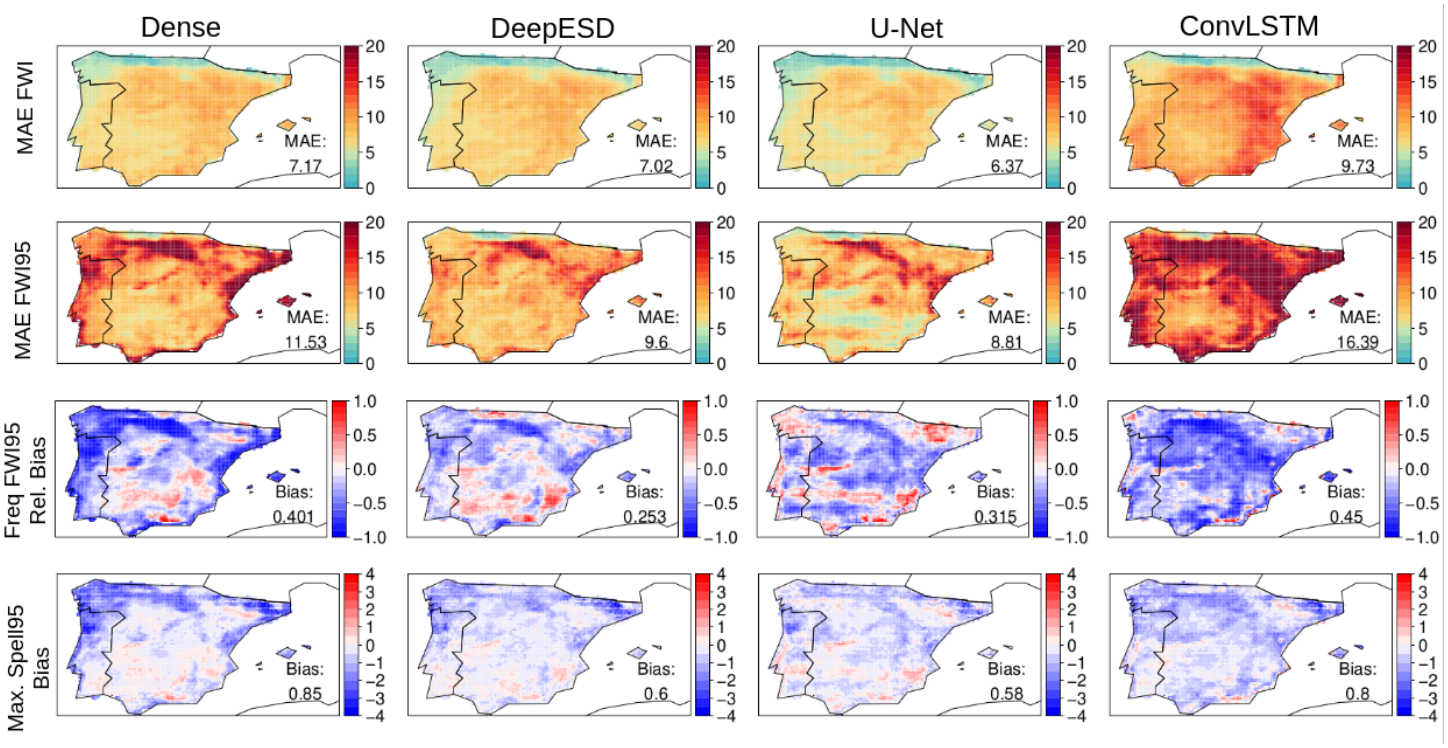

Figure 6: Results from the Dense, DeepESD, U-Net and Conv-LSTM model trained with the P0 predictor set (see Table 1 from manuscript) . The maps display differences relative to the reference FWI for the fire season (June–September) during the test period (2012–2021) for the FWI MAE (first row), FWI95 MAE (second row), Frequency FWI95 Relative Bias (third row) and Maximum annual Spell FWI 95 bias. 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.

In the manuscript we have clarified that we have tested a ConvLSTM, but it is not included in the text in Section 2.4: "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."

<u>Minor 10: Interpretation of input variable relevance</u>

The saliency maps suggest that precipitation is only relevant in low-FWI scenarios. However, this may be a result of how precipitation contributes to the FWI calculation

itself—namely, it offsets the fuel dryness components. Therefore, in high and extreme FWI events (which typically occur during dry periods), the precipitation input often has a value of zero, contributing little additional information for the DL model.

It would be insightful to give more information about why the model changes its focus depending on the region and the type of FWI (non or extreme value). Otherwise, the only information that this result gives us is that for the DL temperature, relative humidity and wind speed are sufficient for obtaining accurate high and extreme FWI events. But, is this true in reality?

Considering this work uses ERA5Land, the predictor variables (T, RH, P and ws) are non independent from each other. In fact, temperature and dew point temperature (needed to compute relative humidity) are variables calculated by the land surface model in ERA5Land, while total precipitation and wind components are forcing variables interpolated from ERA5. Therefore it is likely probable that temperature and relative humidity in ERA5Land reflects the effects of changes in total precipitation and wind speed and, therefore, these two last variables are not so needed by DL models. This is just a guess and it likely won't be the full explanation of your interpretability results... but in any case it would be very valuable to give more information about this or at least mention it if you agree on this limitation.

We thank the referee for this very thoughtful and detailed comment. We agree that the apparent low relevance of precipitation in high- and extreme-FWI scenarios is strongly linked to how precipitation contributes to the FWI itself. Specifically, precipitation primarily influences the fuel moisture codes, and when FWI values are high, these codes already reflect prolonged dry conditions, making the precipitation input frequently zero and thus uninformative for the DL model. We clarify this point in the revised manuscript to ensure that this mechanism is explicitly discussed.

Regarding the reviewer's second point, we also agree that providing more insight into why the model's feature attribution changes across regions and FWI regimes would strengthen the interpretability section. We will expand our discussion by emphasizing two aspects:

FWI definition dependency – The DL model's focus on temperature, RH, and wind in high/extreme cases mirrors the FWI's own reliance on these variables under dry conditions. However, this does not imply that precipitation (or other inputs) is unimportant in reality for fire risk; rather, it reflects the structure and sensitivity of the FWI metric that the DL model is trained to emulate.

Predictor interdependencies in ERA5Land – As the reviewer correctly notes, ERA5Land variables are not independent: RH is derived from temperature and dew point (affected by precipitation indirectly), while precipitation and wind are assimilated forcings. These dependencies likely explain why the DL model can achieve high predictive accuracy even when precipitation and wind receive lower attribution scores. We agree that this introduces a limitation in interpreting the saliency maps, since the apparent dominance of temperature and RH may partly arise from their embedded relationships with other drivers.

We will incorporate this limitation into the revised manuscript and make it clear that the interpretability results should not be read as definitive statements about the physical importance of individual meteorological drivers in real-world fire danger processes. Rather, they highlight how the DL model leverages the structure of the ERA5Land inputs and the FWI formulation itself.

*Therefore, in **Section 3.5** (in the third paragraph of the discussion) we have incorporated the following explanation which summarizes the previous ideas:*

"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 danger; 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 forces. These interdependencies likely contribute to the model's ability to achieve high predictive accuracy even when precipitation and wind receive lower attribution scores."

For all these reasons, we thank the referee for this comment, which enriches the discussion of the interpretability results in the manuscript.

Minor 11:

I miss an experiment in which you assess how the DL models learn to compute the reference FWI using input variables at 12 UTC. Your experiments P1 and P2 address this question to some extent, but the resulting biases could also arise from difficulties relating daily aggregates to FWI 12 UTC data. It may be worth including this experiment to provide insight into where the obtained biases in the DL models may come from.

We thank the referee for this insightful suggestion. In response, we have included an additional sensitivity experiment designed to evaluate the ability of the deep learning models (deepESD, fully connected (dense) networks, and U-Net architecture) to learn the transfer function that defines the Fire Weather Index (FWI) using input variables at 12:00 UTC, consistent with the temporal resolution of the reference index. This new experiment is now included in the revised manuscript in Figure D1 as part of a new Appendix D: Sensitivity analysis of deep learning models and evaluation of other predictors sets.

This experiment complements our previous configurations (P1 and P2) by isolating the effect of temporal aggregation. Specifically, we compare model performance when trained with instantaneous inputs (i.e., values at 12:00 UTC and 24-hour precipitation) versus daily mean inputs. This comparative framework allows us to disentangle the intrinsic biases of each modeling approach from the additional error introduced by using daily-aggregated predictors.

Our results, summarized in Figure 7 below, show 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 experiments P1 and P2 stems from the mismatch in temporal resolution between the predictors and the reference FWI.

Furthermore, we find that the U-Net architecture yields the lowest intrinsic bias in emulating the actual FWI function. It consistently exhibits the smallest biases in both 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 the U-Net architecture may offer enhanced generalization capabilities. Its ability to maintain low bias across both instantaneous and aggregated inputs indicates robustness to temporal variability, which is a key factor in modeling climate indices.

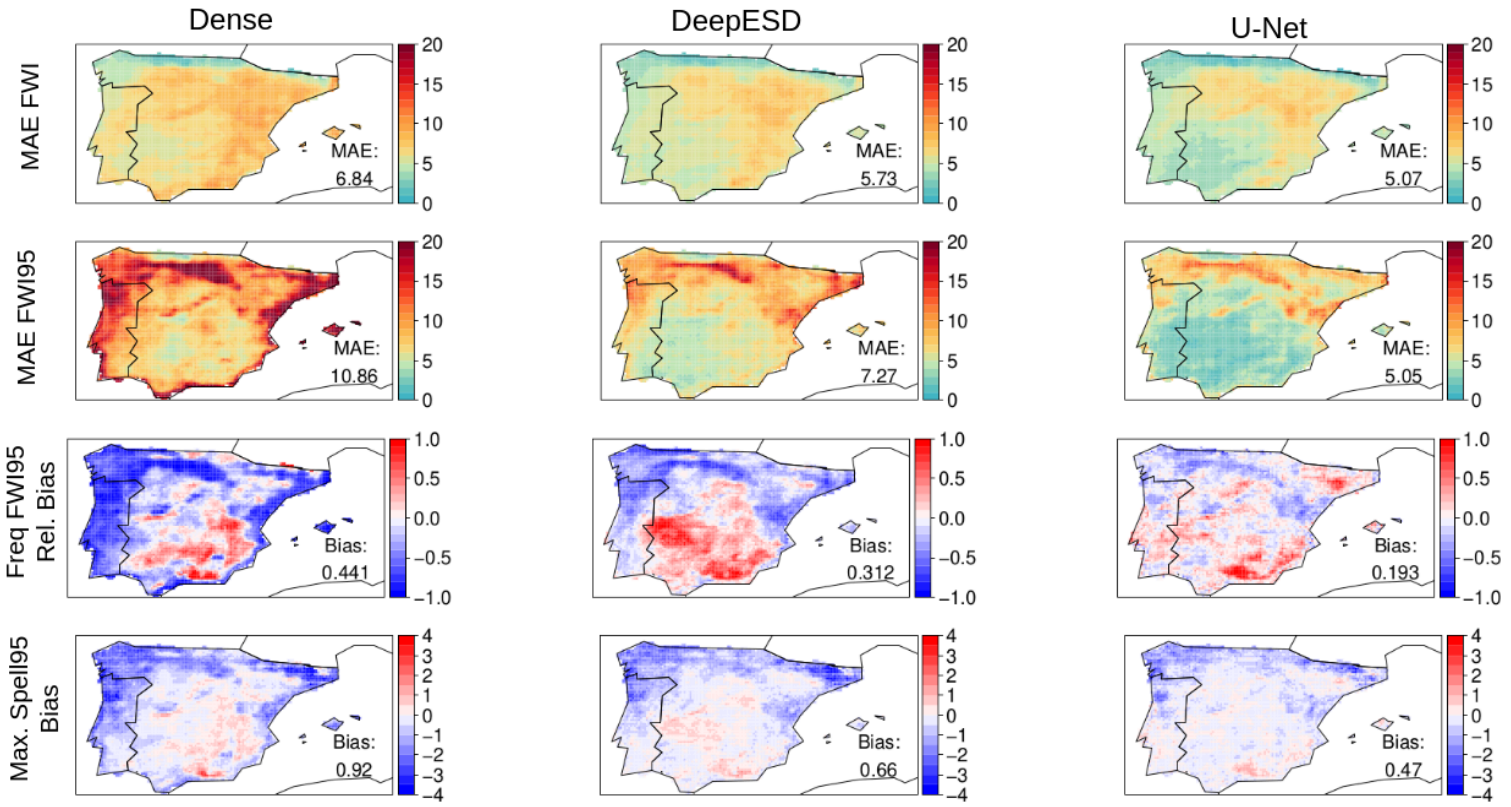

Figure 7: Results from the Dense, DeepESD and U-Net model trained with 12UTC input variables (temperature, 24h-accumulated precipitation, relative humidity and wind speed) . The maps display differences relative to the reference FWI for the fire season (June–September) during the test period for the 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.

A discussion about these results is provided in the new version of the manuscript in Section 3.2.1: "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."