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

I'm a bit confused as to how you have chosen the predictor sets for different experiments, namely P1 and P2. It does not seem to have been done systematically , at least how it is described. For eg. in L115-L120 you say that you have done experiments to avoid precipitation due to it's complex nature, which in itself is fine. But you also say that P1 and P2 are done to improve upon the result of the best proxy P0 which uses DM Temp, Min Rel. Hum., daily accum precip and DM wind speed. The following questions arise based mainly from the way Table 1 is presented:

1. **Why is Min Rel Hum. changed to DM Rel. Hum for P1 and P2 along with removing precip? Were other experiments done with precip included?**
2. **Have you done an experiment with the following as predictors: daily mean Temp, min. Rel. Humidity, and daily mean wind speed?**

In this study, the Deep Learning (DL) models are compared against the optimal Proxy FWI version from Bedia et al. (2014), which serves as our benchmark. This benchmark estimates the daily FWI based on a selection of atmospheric variables: daily mean temperature, minimum relative humidity, daily mean wind speed and 24-hour accumulated precipitation. To ensure the fairest possible comparison between methods, we defined a predictor set for the DL models which contains the latter list of variables (P0).

The subsequent shift from minimum to daily mean relative humidity in predictor sets P1 and P2 is intentional, driven by data availability. Our objective with these new predictor sets was to rely exclusively on daily mean variables, as these are more consistently and widely available in climate model outputs and reanalysis datasets than variables such as minimum relative humidity.

Precipitation presents an additional challenge, as it is a complex variable that climate models often struggle to reproduce. For this reason, we investigated whether robust performance could still be achieved in the absence of precipitation, by excluding it from predictor sets P1 and P2. This design choice also makes the DL models less sensitive to uncertainties in precipitation estimates (L115–L120 in the manuscript). Predictor set P0, which includes daily accumulated precipitation, therefore serves as our baseline configuration. Importantly, our explainability analysis (Section 3.3) confirms that precipitation contributes only minimally to predicting extreme FWI values (e.g., FWI95), reinforcing the notion that the models can maintain strong performance even without it regarding FWI extremes.

Following the reviewer´s question, we have conducted a dedicated experiment using only daily mean temperature, minimum relative humidity, and daily mean wind speed. In this case, we compare P1 with P1 using minimum relative humidity instead of daily mean relative humidity (first and third column of Figure 1). Figure 1 presents the climatology results as is

shown in several Figures in the manuscript: the first row shows the mean FWI, the second row shows the FWI95, the third row indicates the Relative Bias of the FWI95 Frequence and the Maximum Spell95 Bias is presented in the fourth row. As shown, the results obtained with minimum relative humidity are worse in terms of FWI95 MAE than those using mean relative humidity, but similar in terms of mean absolute error (MAE). Moreover, according to the referee suggestions we have tested a U-Net considering the P1 adding precipitation. The results of this aforementioned pattern are competitive with respect to the P1 pattern presented in the manuscript. For instance, P1 + precipitation pattern improves the performance in estimating some of the validation indices assessed. However, we observe that the results for the FWI95 MAE are worse than using the pattern P1. This is linked with the discussion of xAI results in the Section 3.3 of the manuscript. Including precipitation can model better the whole distribution of the FWI since precipitation might be linked to low and moderate FWI events. However, in the case of extremes (FWI95) there is a lack of importance of this variables in the extremes prediction. That may be the reason why U-Net trained with P1 plus precipitation get competitive results in the MAE FWI index, but on the contrary in FWI95 get worse results than the FWI95.

These additional tests have been mentioned in the manuscript in Section 2.3: "*Other predictors sets have been tested, such as P1 using minimum relative humidity instead of daily mean relative humidity or P0 including the lagged 24-hour precipitation, without any added value compared to the configurations mentioned in the manuscript. These experiment results are shown in Figure D2." Moreover, a Figure (Figure D2) has been included in the new Appendix D: Appendix D: Sensitivity analysis of deep learning models and evaluation of other predictors sets.*

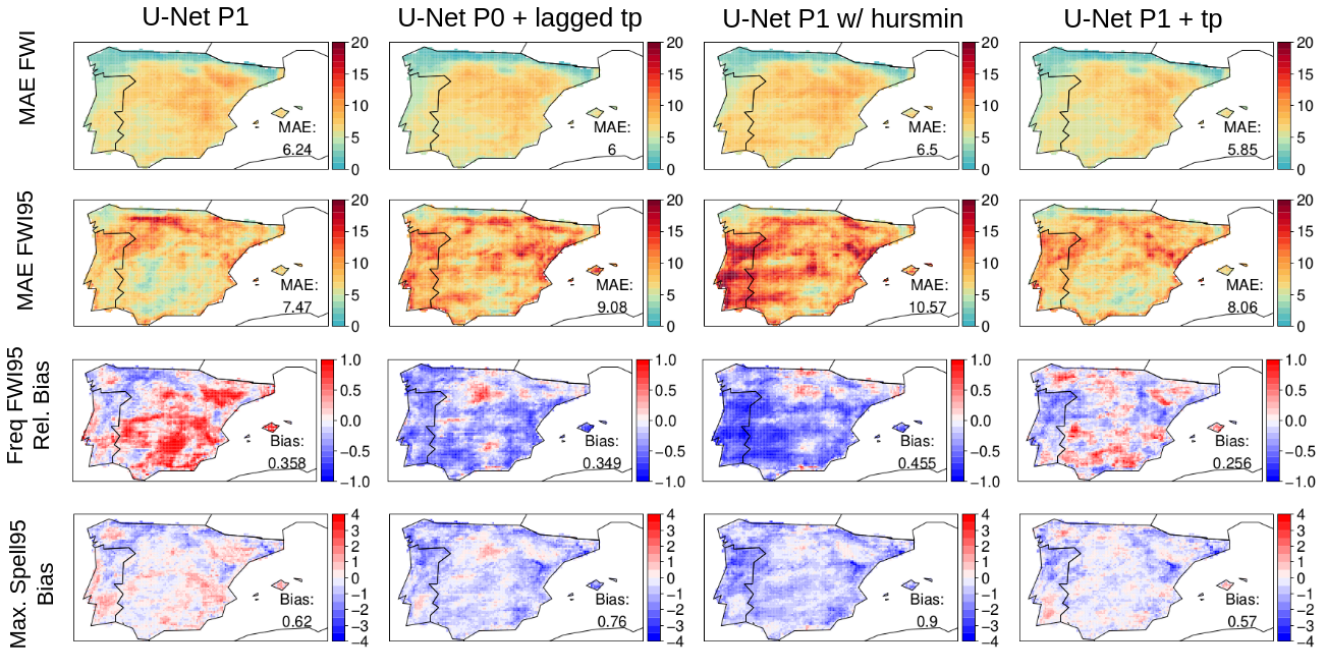


Figure 1: *From left to right, results from the U-Net model trained with the P1 predictors set (see Table 1 from manuscript), trained with P0 and adding the 24-h lagged precipitation,*

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 (2018–2021) 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.

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

It may actually be only required to perhaps restructure L115-L120 to better suit Table 1 to avoid confusion.

We acknowledge that lines L115–L120 in the original manuscript could create confusion for the reader regarding the motivations behind the predictor configurations. In particular, the sentence “[P1 and P2 are evaluated as alternative emulation inputs (Table 1), aiming to improve upon P0 results.]” was unclear and potentially misleading. To address this issue, we have revised the paragraph in the updated version of the manuscript to clarify our rationale and improve readability.

“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 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.”

L155-L165: How do you make sure that the model doesn’t overestimate low-moderate values to compensate for extreme value underestimation? Was this tested at all? I

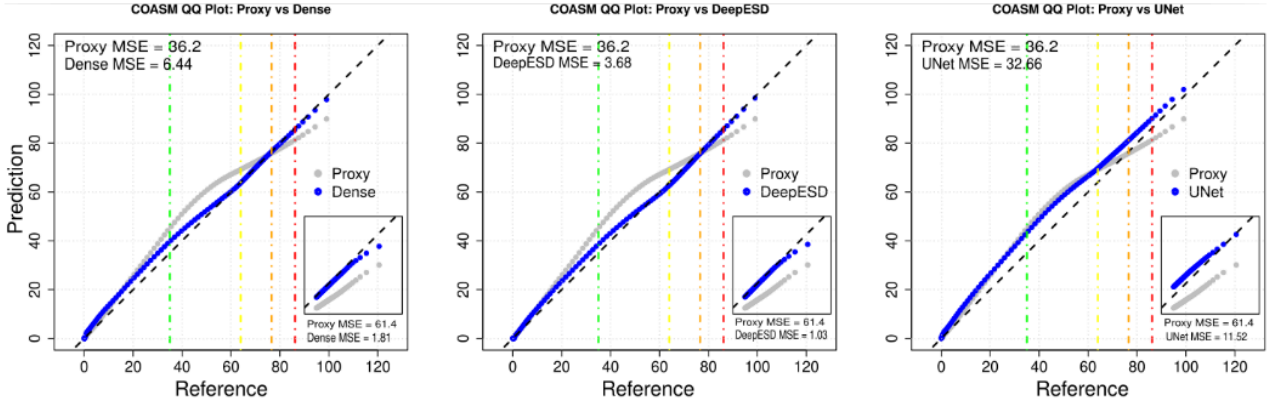
understand that it may not be relevant if you're only interested in extreme cases. However, the total performance of the model still needs to be physically consistent.

Thank you for your comment. The ASYM loss function is specifically designed to better capture extreme events, particularly in the right tail of the distribution. We have verified that models trained with this loss function do not overestimate low to moderate values in the case of the U-Net model. However, for the other models (Dense and DeepESD), we observe an overestimation in the lower and moderate ranges. A graphical comparison is provided below, showing the distribution of predictions using QQ plots for a deep learning model trained with both the MSE and ASYM loss functions.

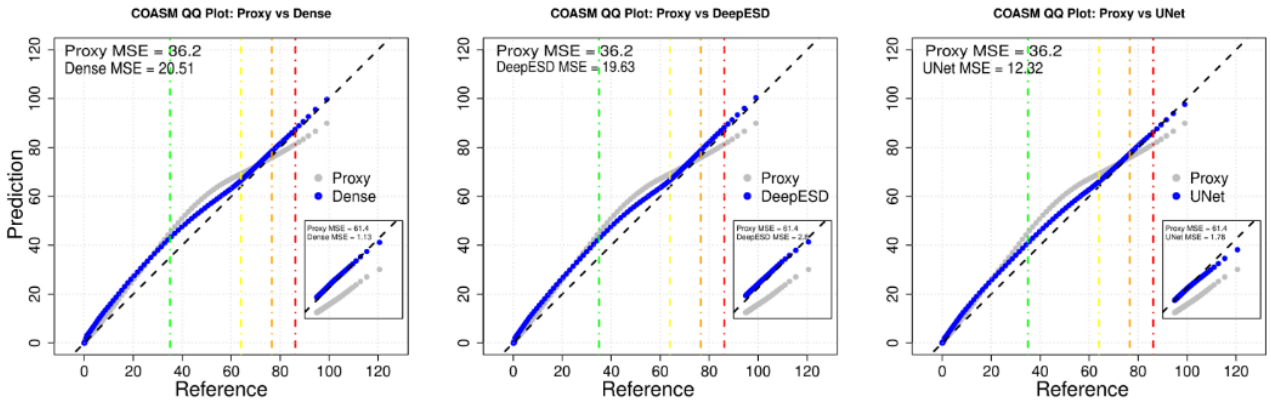
QQ plots were constructed to assess the agreement between observed FWI, proxy FWI, and deep learning model predictions. For each model, empirical quantiles of the observed FWI were computed at probability levels from 0 to 0.99 in steps of 0.01. The same quantiles were calculated for both the proxy and model outputs, using the FWI daily values for all grid points concatenated after spatial subsetting and seasonal masking.

The QQ plots were then obtained by plotting the observed FWI quantiles (reference distribution) against the corresponding quantiles of the proxy and the model. In addition, mean squared errors (MSE) between the observed FWI and proxy quantiles, as well as between the observed FWI and model quantiles, were computed to quantify discrepancies. To further evaluate model performance in the upper tail of the distribution, the procedure was repeated for the 95th to 99.9th percentiles (in steps of 0.001), and the resulting QQ plots were shown as inset panels. Finally, threshold lines corresponding to key percentiles (40th, 65th, 85th, and 95th) of the observed distribution were included to highlight different fire risk categories.

MSE



ASYM



the models architectures (Dense, DeepESD, U-Net). The solid blue line represents the model predictions, while the black dashed line indicates the 1:1 correspondence. Vertical dashed lines mark percentiles of the proxy distribution.

Figure 2 presents Q-Q plots comparing predicted and proxy values for different models trained with two different loss functions: MSE (first row) and ASYM (second row). A key inspection of these plots show us that the Dense and DeepESD models with ASYM as loss function overestimates low and moderate values of the distribution, a behavior which is not observed when training with the MSE. In the case of the U-Net model this overestimation is not observed, and indeed training with the MSE as loss function yields a slight overestimation of low and moderate values. This suggests that the ASYM loss function—while tailored to better capture extreme events—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 less accurate than when considering the MSE as loss function. Therefore, the ASYM loss function can compromise the performance of low and moderate levels in some cases. This insight has been added to the manuscript in Section 3.5: "These findings suggest that while ASYM is generally advantageous for modeling the tail of the distribution, it can sometimes lead to an overestimation of low and moderate FWI values. While the ASYM loss function does not compromise the U-Net model performance in the lower quantiles, it leads to an overestimation for the Dense and DeepESD models of low and moderate values for the FWI as compared to results from models trained with MSE as loss function. Therefore, the ASYM loss function can deteriorate 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." Also we want to highlight that the QQ-Plot Figure is added as Figure C1 in the new appendix section C: QQ-plots assessment.

L220: Is this a description of Fig. 2?

Yes, line 220 refers to Figure 2. Thank you for highlighting this typo.

L256-L261: The authors may perhaps explain if the smoother outputs from DL models are good/bad? And perhaps also give an explanation why it's smooth.

1. **Have you considered using some form of terrain-based predictor, say topography field for example? It might help in providing a more discernible profile to the fields.**
2. **Since you're estimating FWI, I would expect that vegetation cover would be a necessary predictor. Why was it not used?**

We thank the referee for the thoughtful observations. Regarding the smoothness of the DL model outputs, we acknowledge that this is a common feature of neural network-based emulators, particularly when trained to minimize mean squared error. In our view, the smoother fields are not necessarily detrimental; rather, they reflect the models' tendency to average out local noise and variability that may not be strongly represented in the training data. In operational contexts, smoother outputs may even be advantageous, as they reduce false alarms and improve spatial coherence. Nonetheless, we agree that further work could

explore methods to enhance spatial detail, such as Super-Resolution Techniques, spatial attention mechanisms or other post-processing techniques.

Concerning the use of terrain-based predictors, we appreciate the suggestion. While topography can influence fire behavior and fuel moisture indirectly, the Fire Weather Index (FWI) is designed to quantify meteorological fire danger independently of terrain or vegetation characteristics. It is computed solely from weather variables (temperature, humidity, wind, and precipitation), and does not include any static environmental inputs. Therefore, introducing terrain predictors would not be consistent with the formulation of the reference index we aim to emulate.

Similarly, vegetation cover is not used in the computation of the FWI. The index is not a fire risk or impact model, but rather a meteorologically driven indicator of fire danger potential. It assumes a generic fuel type and does not account for land cover, fuel load, or vegetation type. For applications where vegetation is relevant—such as fire spread modeling or impact assessment—additional predictors would indeed be necessary. However, in the context of this study, which focuses on emulating the FWI as defined by its original formulation, the inclusion of vegetation data is not required.

L300-L302: If it is not too much extra work, I would like to see how previous days' (e.g. 24h prior) precipitation (lagged precipitation) might affect the model capabilities? The simplest model would suffice. If it is not possible to run the models, then you can also explain how this would affect/not affect the models' performance (with necessary refs).

We appreciate the suggestion to include lagged precipitation as an additional predictor. To evaluate its impact, we trained a version of the U-Net model using the P0 set (surface air temperature, minimum relative humidity and surface wind speed), adding 24-hour lagged precipitation as an extra input.

As shown in the result , the inclusion of lagged precipitation in the predictor set leads to noticeable changes in model performance (see Figure 1 of this document, columns 1-2). While it helps to reduce certain biases, this improvement comes at the cost of higher errors in the upper tail of the FWI distribution, particularly for extreme values. This trade-off suggests that lagged precipitation does not consistently enhance the model's ability to capture critical fire danger conditions, which are most relevant for our analysis.

These results are consistent with the saliency analysis presented in Section 3.3 (Figure 5 in the manuscript). As shown in the manuscript, precipitation gains importance when predicting low to moderate FWI values; however, for extreme FWI levels (high, very high, or extreme), the contribution of precipitation is negligible. This could explain why the MAE across the full FWI distribution is lower when lagged precipitation is included as an input, while performance for FWI95 worsens under the same setting. Since our main focus is on

accurately emulating extreme FWI values in the right tail of the distribution, we consider that including lagged precipitation in the predictor set is not the most suitable approach. These additional tests have been mentioned in the manuscript in Section 2.3: "Other predictors sets have been tested, such as P1 using minimum relative humidity instead of daily mean relative humidity or P0 including the lagged 24-hour precipitation, without added value compared to the configurations mentioned in the manuscript. Therefore, these experiments are not included in the text for simplicity"

It is also important to note that the FWI system itself embeds memory of past wet and dry conditions. Its subcomponents, particularly the Drought Code (DC) and Duff Moisture Code (DMC), accumulate the effects of precipitation and drying over multiple days. Consequently, even without explicitly including lagged precipitation as an input, the target variable (FWI) inherently reflects the influence of prior weather conditions. This allows the model to learn these dependencies implicitly from the current meteorological inputs.