Review of Deep Learning Emulation of Multivariate Climate Indices: A Case Study of the Fire Weather Index in the Iberian Peninsula. Rev3 RC3

The manuscript has improved after revision, and the additional analysis is useful and relevant. The authors have addressed most of my previous comments satisfactorily. However, I believe a few minor points still deserve clarification before publication.

Regarding my initial point 2 : "Another potentially misleading aspect is the use of the original FWI variables in its computation, compared with the predictors used in the emulation models. It would strengthen the manuscript if the authors explicitly mentioned this distinction, either within the main text or at least in the conclusions section"
My original comment was mainly about explicitly acknowledging the conceptual distinction between:
(i) the reference FWI computed directly from the original meteorological variables entering the standard FWI formulation, and
(ii) the predictors used by the DL emulators, which may differ due to temporal aggregation and/or preprocessing.

I still encourage the authors to state this distinction explicitly in the manuscript (e.g. in the methodology, or conclusions), since it is important for properly interpreting the reported emulator errors and avoiding potentially misleading comparisons. Without this clarification, the results may be misleading or overly interpreted.

We sincerely thank the reviewer for their careful reassessment of the revised manuscript. We are pleased that the reviewer acknowledges that the manuscript has improved after revision and that we have addressed most of the previous comments satisfactorily. We particularly appreciate their constructive guidance on the remaining point that requires further clarification.

Upon reflection, we recognize that our previous revision did not make the distinction between reference FWI variables and DL predictor sets sufficiently explicit and prominent to adequately address the reviewer's concerns. The reviewer is absolutely correct that this is a critical issue for properly interpreting the emulator errors and avoiding potentially misleading comparisons, and we have therefore prepared substantially more explicit clarifications that are now integrated throughout the manuscript.

The distinction the reviewer has identified is indeed fundamental to understanding our work. The reference FWI is computed directly from instantaneous meteorological variables measured at noon local standard time (LST), as originally defined in the Canadian FWI System. Our DL emulators, however, are trained using daily-aggregated versions of these same variables (daily means of temperature, wind speed and relative humidity)rather than instantaneous noon values. This is not a minor technical detail but rather a central characteristic of the emulation task that must be clearly understood by readers to avoid misinterpreting our results. We ensure this distinction is now stated explicitly, prominently, and repeatedly throughout the manuscript to make clear this point.

Specifically, we add a new paragraph at the very beginning of Section 2.3 "Predictor Sets" in the Methodology section that explicitly establishes this conceptual distinction before any technical details are presented. This opening paragraph clearly states that the reference FWI requires instantaneous noon LST variables while the DL predictor sets consist of daily-aggregated variables, and explains that this fundamental difference in temporal resolution represents the core characteristic of the emulation task.

We further emphasize this distinction through a minor change in the caption of Table 1, which summarizes the predictor sets: “The 'FWI' row indicates the instantaneous noon local standard time (LST) variables required by the standard FWI definition. The predictor sets P0, P1, P2 represent temporal aggregations or modifications of the noon LST variables, such as daily means, which are far more widely available across reanalysis and climate datasets. This motivates the emulation approach: rather than requiring instantaneous noon inputs, the emulators allow FWI to be derived directly from these more accessible variable forms.”

In the Conclusions section, we add a substantial new paragraph dedicated specifically to this distinction, at the beginning of the section: “A key methodological distinction should be emphasized for the proper interpretation of these results: the reference FWI is computed from instantaneous noon LST meteorological variables as originally defined in the FWI system, whereas the DL emulators are trained using daily-aggregated predictors (daily means, minima, or accumulated values). The reported emulator skill should therefore be understood as the ability of DL models to recover a noon-time fire-weather signal from daily-aggregated inputs, a practically valuable capability precisely because such inputs are far more widely available than instantaneous noon observations.”The paragraph explicitly states that our results demonstrate a solution to a real data-limitation problem in climate applications, not evidence that the temporal requirements of the FWI system is less important than the original designers understood.

These revisions ensure that the distinction between reference FWI variables and DL predictor sets is stated explicitly, prominently, and repeatedly throughout the manuscript.. We hope that the revisions have sufficiently addressed the reviewer’s concerns.

Conclusions section:
I also encourage the authors to discuss more carefully the interpretation of the precipitation-free experiments. In the standard FWI formulation, precipitation is a key input variable affecting several moisture codes/sub-indices (e.g. FFMC, DMC, and particularly DC). Therefore, the fact that the emulator can reproduce the final FWI with limited skill degradation when precipitation is omitted from the predictors does not necessarily imply that precipitation is physically unimportant for fire-weather conditions. Rather, the DL model may be indirectly recovering part of this information through correlations with other variables and/or through the characteristics of the JJAS climatology.

I would like to stress that I do not have concerns regarding the scientific approach or the validity of the emulator framework presented in the paper. However, I am slightly concerned that some readers may interpret the emulator results from a physical perspective, particularly regarding the experiments excluding precipitation. I therefore believe that this distinction should be clarified more explicitly in the manuscript, particularly in the conclusions section, to avoid potential misunderstandings or over-interpretation of the results.

We thank the referee for this thoughtful comment, which highlights an important distinction between model behavior and physical interpretation. The reviewer's concern is well-founded and reflects a nuance that we should have emphasized more explicitly in our original manuscript.

We fully agree with the reviewer that the ability of the emulator to reproduce the FWI with limited skill degradation when precipitation is omitted does not imply that precipitation is physically unimportant for fire-weather conditions. Rather, as the reviewer notes, the deep learning model may indirectly recover information about precipitation effects through complex linear and nonlinear statistical dependencies with other meteorological variables (such as temperature, humidity, wind, which are likely to covary with precipitation) and/or through learned patterns of the JJAS climatology.

We have revised the conclusions section to clarify this distinction more explicitly. Specifically, we now emphasize that the precipitation-free experiments reveal a property of the emulator's learned representations, namely that other variables contain sufficient information to approximate the FWI signal under JJAS conditions, rather than a statement about the fundamental physical role of precipitation in the FWI system. We have also added a brief discussion of the potential mechanisms (indirect information recovery through covariability) by which the model achieves this performance.

We hope that these revisions ensure that readers understand the limitations of this finding when interpreting results in a physical context, while still acknowledging the practical utility of such precipitation-free emulators for applications where precipitation data are unavailable or uncertain.

Review of Deep Learning Emulation of Multivariate Climate Indices: A Case Study of the Fire Weather Index in the Iberian Peninsula. Rev3 RC4

Dear authors
Thank you for your answers and the effort done to improve the quality of the paper and to respond to main caveats highlighted.
However some of my crucial doubts were not completely clarified, namely:

Dear Professor Gouveia, first of all we would like to thank you for your continued engagement with the manuscript and for the detailed comments provided in this round. We appreciate the concerns raised, that have helped us to reflect on the assumptions and main message of the article to avoid misunderstandings. We have also carefully revisited the corresponding sections of the manuscript to further clarify the scope, assumptions, and interpretation of our results.

While we acknowledge that the contribution may not fully align with the reviewer's perspective, we believe that the proposed approach offers a valuable alternative to the use of proxy variables for FWI calculation in climate-scale applications, as it is often done in the literature due to data availability constraints. In this context, our results consistently show that the proposed emulators provide a more accurate representation of both mean conditions and extreme events, suggesting their potential to improve current practices in many scientific studies using proxy FWI versions. In addition, we incorporate interpretability into the deep learning models, moving beyond a purely "black-box" approach towards a more physically grounded understanding of their behavior.

- The authors said: "The practical value of the proposed emulators lies primarily in their capacity to generate rapid, spatially consistent fire-weather information suitable for climate services, climate-change impact assessments, and large-scale evaluations of natural hazards."
In my opinion, to achieve this goal, the accuracy and usability of FWI should be the same as for other operational purposes; otherwise, the index does not represent the reality. If the authors think another index should be better, they should propose it and validate it.

Therefore, if the aim is to propose an AI method that emulates FWI, focusing in the AI techniques and consistency measures, I would suggest publishing the work in a journal focused on AI

We thank the reviewer for this comment and for raising an important point regarding the intended application of the proposed approach.
We would like to clarify that the objective of this study is not to achieve the level of accuracy required for real-time operational fire management or decision support for wildfire suppression. Rather, the term "operational" was intended in a broader climate-services context, referring to the ability to efficiently generate spatially consistent fire-weather information for climate monitoring, impact assessments, and large-scale hazard analyses. We acknowledge that this distinction may not have been sufficiently clear in the original manuscript and have revised the text accordingly.
In many climate-scale applications, including those based on GCM or RCM outputs, the instantaneous meteorological variables required for direct FWI computation are not readily available. Under these circumstances, it is common practice in the literature to rely on

proxy-based approximations of FWI. However, as demonstrated in our results, such proxy approaches can lead to sub-optimal representations of both mean conditions and extreme events.
The contribution of this work is therefore to propose and rigorously evaluate an alternative methodology, based on deep learning emulators, that better reproduces the behavior of the reference FWI under these constraints. Importantly, our goal is not to redefine the index itself, nor to replace it, but to improve its estimation in contexts where its direct computation is not feasible.
From this perspective, we consider that the manuscript falls well within the scope of NHESS, as it presents the design, evaluation, and validation of a methodological framework aimed at improving the modelling of a widely used natural hazard indicator under climate-change conditions. We have revised the manuscript to better emphasize this scope and to avoid any ambiguity regarding the intended applications of the proposed method.

- The author said: "At the same time, we would like to clarify that the objective of the present study is not to reproduce the full operational structure of the CFFDRS, nor to provide a forecasting tool for real-time wildfire suppression activities. As now emphasized in the revised manuscript, our goal is specifically to emulate the reference FWI, which is the most widely used indicator in climate-scale applications, impact assessments, and large-area evaluations of natural hazards."
Considering the argumentation and the provided results DC, DMC and FFMC, which are not included in the manuscript, I am not happy with the results obtained with DC and DMC. In particular, in the case of DC (figures 3 and 4), the temporal evolution has a high variability do not consistent with the reference data and the definition of DC. This result is probably due to the non-inclusion of precipitation.

We respectfully clarify that the primary objective of this study is not to redefine or replace the Fire Weather Index (FWI), nor to provide an operational forecasting tool, but rather to evaluate the capability of deep learning models to emulate the reference FWI within a climate-oriented framework. In this sense, the goal is to reproduce the behavior of an established and widely used index under controlled experimental conditions, rather than to propose an alternative index.
We agree that, in operational contexts, accuracy requirements may be more stringent. However, in climate-scale applications (such as impact assessments or large-scale hazard evaluation) the emphasis is often placed on the robust representation of spatial patterns, temporal variability, and change signals, rather than on exact replication at the event level. We have clarified this distinction more explicitly in the revised manuscript.

We also acknowledge the reviewer's comments regarding the sensitivity of DC and DMC to precipitation. As correctly noted, these components are influenced by rainfall events, particularly when thresholds are exceeded. In our framework, precipitation is not included as an explicit predictor, and we agree that this may affect the representation of short-term variations in these indices.

However, the aim of the study is to assess the ability of the models to reproduce the aggregate behavior and long-term variability of the FWI system from a reduced predictor set. In this context, the results should be interpreted in terms of overall signal representation rather than exact event-level fidelity. We now explicitly acknowledge this limitation in the manuscript and clarify that the representation of sub-daily or precipitation-driven variability may be partially smoothed in the emulator outputs.

These clarifications have been incorporated in the revised manuscript, particularly in the Introduction and Conclusions, where the intended climate-services context of the proposed approach is now explicitly defined.

- The author said “On the other hand, as the "long-term memory" of the system, the DC is only affected by significant or sustained rainfall that can penetrate deep, compact organic layers. This means that even during severe events, light may temporarily lower surface ignition risk (FFMC) but fails to penetrate deeper fuel layers”
I do not agree, as per definition of DC a value higher than 2.8 mm as an impact in DC discontinuing the normal increase of DC during summer. Please see Van Wagner and Pickett (1985), Van Wagner (1987) and Lawson and Armitage (2008).
Please see Figure 6 of Ramos et al. (2023), that show a drop in DC associated with precipitation occurrence and the example of results of the table below:

Day	Temp (°C)	Precip (mm)	Drought Code (DC)	Trend Note
1–10	~26°C	0.0	300.0 → 353.4	Steady drying (approx +5.3/day)
11	23.6°C	15.0	308.9	Major Drop (-44.5 units)
12	23.6°C	5.0	302.3	Further drop from follow-up rain
13–20	~23°C	0.0	308.1 → 344.2	Resumed drying phase
21	29.4°C	2.0	350.7	DC increases (Rain < 2.8mm ignored)
22–30	~24°C	0.0	356.3 → 400.0	Reaching high seasonal drought levels

We thank the reviewer for this comment and for highlighting the formal definition of the Drought Code (DC) response to precipitation.
We are aware that, according to the standard formulation of the FWI system, precipitation above a threshold of approximately 2.8 mm can affect the evolution of the DC by interrupting or reversing its increasing trend. Our intention in the cited paragraph was to provide a conceptual interpretation of the DC as a component with a longer memory compared to the surface indices (fine fuels and duff moisture codes), emphasizing its slower response to precipitation and its association with deeper fuel layers. However, we acknowledge that the wording may have suggested that only large or sustained precipitation events influence DC, which is not strictly consistent with its formal definition. We have revised the corresponding description of the DC in Section 2 to better align it with the formal definition of the FWI system and avoid potential ambiguity.

- The authors argue: “While a detailed event-by-event analysis of historical wildfire episodes lies beyond the scope of the present study (which focuses on evaluating the performance of deep learning emulators trained on daily predictors) we have incorporated several complementary analyses that directly address the reviewer’s concerns”
In fact, without the analysis of extreme cases proposed or a statistical analysis of extremes, I keep my opinion about the performance of the method.

We thank the reviewer for reiterating this point and fully agree that the representation of extreme fire-weather conditions is a key aspect for many applications.

However, we would like to clarify that the present study is not designed as an event-based evaluation of individual wildfire episodes, but rather as a systematic assessment of model performance at the regional and climate scales. In this context, our evaluation focuses on the ability of the proposed methods to reproduce the statistical and spatial characteristics of FWI, including its behavior at high values.

To address this aspect, the manuscript already includes complementary analyses that explicitly target extremes, including the evaluation of categorical exceedances and the representation of high-percentile values. These approaches provide a consistent and statistically robust assessment of extreme conditions at the temporal and spatial scales considered, which are more appropriate for climate-scale applications than individual case studies.

We acknowledge that a detailed event-by-event analysis could provide additional insights, particularly for operational or case-specific investigations. However, such analyses fall outside the scope of the present work, which is focused on the general performance and robustness of deep learning emulators for climate-oriented applications. We have also revised Section 3 (Results and Discussion) and the Conclusions to more clearly highlight the statistical evaluation of extremes already included in the manuscript.

Therefore, I consider that my main concerns remain, and I can not agree with the publication of the present manuscript.

While we acknowledge that the reviewer's concerns remain, we respectfully note that they largely reflect a different perspective on the scope and objectives of the study. In our view, the proposed approach, focused on the emulation of FWI behavior under data constraints, provides a valuable contribution for climate-scale applications, where proxy-based approximations are still widely used. We believe that this contribution is well aligned with the aims and scope of NHESS.

We thank the reviewer again for the detailed comments, which have helped us improve the clarity of the manuscript, particularly regarding its scope, assumptions, and limitations. We hope that the revised version now provides a clearer and more precise framework for the interpretation of our results.