We answer here to the comments made by Referee 2

We divide the comments, as done by Referee 2, into the Main Comments and Specific Comments sections, with the respective comments.

Main Comments

The heterogeneity in CO2 SHAP contributions across locations (Fig. 5) is puzzling. If CO2 is capturing the anthropogenic warming signal, one would expect relatively homogeneous contributions, perhaps modulated by maritime/continental influences. Instead, values range from 5.6% (Hannover) to 21.8% (Córdoba), and this pattern does not obviously map onto any physical explanation. I suspect CO2 may be acting as a proxy for the location-specific trend in extreme frequency (cf. Fig. 4) rather than cleanly representing anthropogenic forcing. This warrants discussion and suggests non-negligible methodological uncertainty.

We appreciate the reviewer's comment regarding the heterogeneity of CO2 SHAP contributions. While CO2 is a global anthropogenic forcing, its manifestation in local weather —and thus its 'importance' in a statistical model—is not expected to be spatially uniform. This aligns with the findings of Sippel et al. (2020), who show that the fingerprint of climate change is now detectable in daily weather patterns, but this signal manifests with significant spatial variability.

Therefore, the differences in CO2 importance among our locations reflect this spatially non-uniform emergence of the climate signal. We have updated the manuscript to clarify this point in L270 of section 3.2:

These differences among locations align with the work of Sippel et al. (2020)in that different regions show different responses to climate change, with the forced fingerprint manifesting through varying signal-to-noise ratios across the globe.

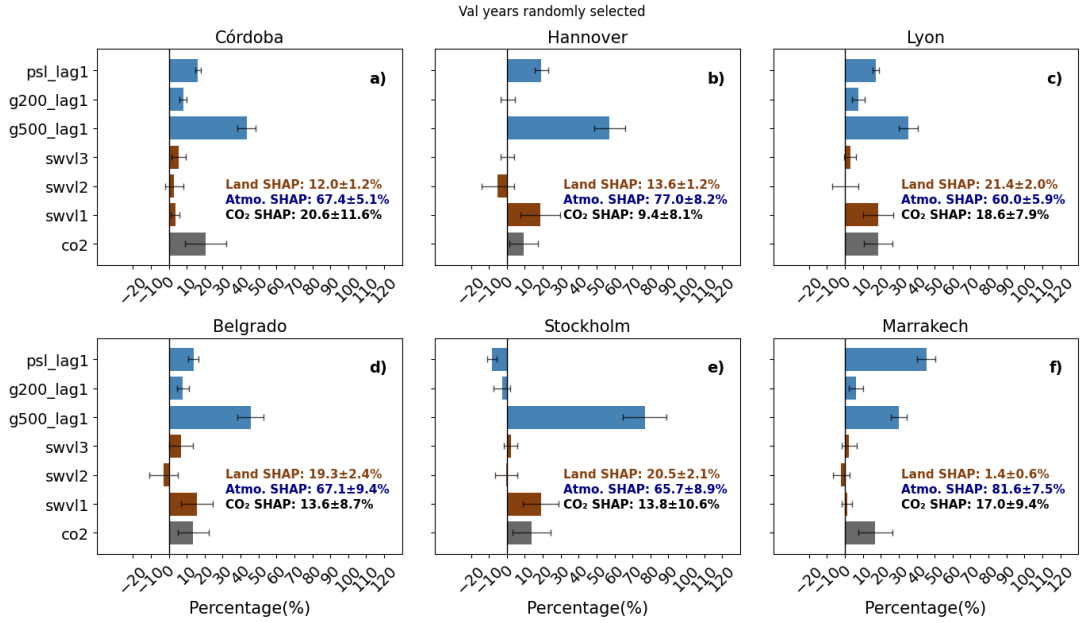
Related to this: given the 20-member ensemble, it should be straightforward to show uncertainties on the SHAP values. This is notably absent and would substantially strengthen confidence in the reported differences between locations and features.

We thank the reviewer for this suggestion. We agree that showing the ensemble spread is crucial for assessing the robustness of the feature importance. Accordingly, we have updated Figure 5 to include uncertainty bars for the SHAP values.

These bars represent the Standard Error of the Mean (SEM), calculated by taking the standard deviation of the SHAP values across the 20 ensemble members and dividing by the square root of the ensemble size. This provides a clear measure of the stability of our results across different model initializations.

For the remaining figures, we chose not to include uncertainty bars in order to preserve the readability of the plots, as adding uncertainty bars would substantially reduce visual clarity. We have clarified this in the revised manuscript in the description of Fig. 5, which now reads:

Figure 5. Extreme class prediction mean SHAP value percentage for all locations using the CombinedModel, taking the 50% most confident predictions. The x-axis represents the percentage contribution of each feature relative to the total sum of mean absolute SHAP values. Error bars indicate the Standard Error of the Mean (SEM) derived from the 20-member ensemble, representing the uncertainty across different model initializations. The colored text box provides the aggregated percentage contributions for Land and Atmospheric components, as well as the CO2 contribution, with associated ensemble uncertainties (±SEM).



The claims regarding physical understanding (e.g., L. 352: "These insights enhance our understanding of the physical mechanisms behind temperature extremes") should be tempered. The role of atmospheric circulation and land–atmosphere feedbacks in modulating heat extremes is well-established from both observations (e.g., Hirschi et al., 2011) and model experiments spanning at least two decades (e.g., Schär et al., 1999; Koster et al., 2004; Fischer et al., 2007; Wehrli et al., 2019; Wehrli et al., 2022). The contribution here is methodological (demonstrating the applicability of XAI techniques to this problem) rather than providing fundamentally new physical insight. The paper would benefit from framing it as such, and from engaging more thoroughly with existing literature (of which I have only cited some examples).

We agree with the reviewer that the physical drivers of temperature extremes—specifically the roles of atmospheric circulation and land–atmosphere feedbacks—are well-established in the literature. Our intention was not to claim the discovery of new physical mechanisms, but rather to demonstrate that our XAI framework can effectively quantify these contributions in a manner consistent with established theory.
We suggest re-writting the final sentence of the abstract (as the text in L352 was mistakenly repeated from the Abstract) to:

These results align with the established physical roles of atmospheric circulation and soil moisture modulation for temperature extremes, and demonstrate the methodological potential of explainable artificial intelligence to quantify the relative contributions of these drivers in a data-driven framework.

We have re-written other parts of the text that also should be tempered and taken some of the references proposed by the Referee.

- L396: The correlation between SHAP values and standardized soil moisture anomalies is negative across all levels, confirming the role of dry conditions in enhancing the probability of extreme heat events
    - New: *The correlation between SHAP values and standardized soil moisture anomalies is negative across all levels, consistent with the well-documented role of dry conditions in enhancing the probability of extreme heat events (Hirschi et al., 2011; Seneviratne et al., 2006).*

- L417: These results reveal the central role of atmospheric circulation in driving extreme temperature events and highlight the regionally dependent modulation by soil moisture.
    - New: *These results reproduce the well-known central role of atmospheric circulation in driving extreme temperature events and highlight the regionally dependent modulation by soil moisture (Fischer et al., 2007; Barriopedro et al., 2023).*

- L447: By combining a robust machine learning method with an ML explainability technique, we offer insights that can inform both model development and climate attribution efforts.
    - New: By combining a robust machine learning method with an ML explainability technique, we provide a framework that can support model development and climate attribution efforts

- L276: …further confirming the robustness of the identified land-drying feedback mechanism across different methodological choices.
    - …further confirming the robustness of the land-drying feedback mechanism identified by the model across different methodological choices.

On a related note, some claims would be more compelling if results were shown primarily for "pure" observations. Since ERA5-Land is an offline land surface model simulation, its behaviour bears similarity to land components of ESMs that have long pointed to these feedbacks. The robustness tests using E-OBS and SPEI are appreciated, but greater emphasis on observational constraints would strengthen the analysis.

We propose including Fig. D2 in the main text and extending the discussion related to the XAI results using EOBS data. Additionally, we suggest including an extended discussion on the need to use such proxies if this type of methodology is to be applied to climate models such as CMIP6. Further work that we aim to do with this methodology is comparing the physical consistency of CMIP6 models in quantifying the role of the different drivers. The treatment of temperature and precipitation among CMIP models is more consistent than that of soil moisture and land variables, and for this reason, the SPEI proxy would be used for the comparison task.

Specific Comments

L.69: The term "arid" is imprecise here. Hsu & Dirmeyer (2023) distinguish "dry," "transitional," and "wet" soil moisture–evaporation regimes; Córdoba would fall into a dry/transitional regime rather than being truly arid climatologically.

Thank you for pointing this out. We agree on changing the terminology and including the proposed reference when describing the regimes for both locations. The text now would read:

These events were selected to represent contrasting soil-moisture feedback regimes – dry/transitional versus transitional/wet – with differing soil moisture variability and land-atmosphere coupling strengths (Hsu and Dirmeyer, 2023).

Hersbach et al. (2020) should be cited for ERA5.

Thank you for your suggestion. We now include this reference and the one for Soci et al. 2024.

L. 87 vs. L. 116: The justification for using ERA5-Land based on resolution is unconvincing given that the analysis ultimately operates at 1° resolution with 100 km spatial averaging. The more relevant advantage is that ERA5-Land, as an offline (land surface model) simulation, does not assimilate screen-level observations to adjust soil moisture (unlike ERA5), meaning soil moisture variability is more physically consistent.

Thank you for the suggestion. We agree on the reasoning offered by the referee and we clarify it in the text now:

Soil moisture variability is more consistent in ERA5-Land because it does assimilate screen-level observations to adjust soil moisture (unlike the ERA5 reanalysis) (Muñoz-Sabater et al., 2021).

Section 2.2: It would help to briefly explain upfront why SPI/SPEI are introduced, since the main analysis relies heavily on ERA5-Land and the drought indices only appear much later.

Thank you for the suggestion. We agree that this might be helpful for the readers. Section 2.2 can be updated to explain that:

1. **Validation & Robustness:** These indices serve as proxies for checking the robustness of our ERA5-Land results against observational data. Because long-term, spatially continuous soil moisture observations are limited, SPI and SPEI (derived from observational precipitation and temperature) provide a reliable alternative for historical comparison.
2. **Model Consistency (CMIP6):** A key objective of this methodology is its future application to CMIP6 models to assess the physical consistency of these models in representing the drivers of hot temperature extremes. While soil moisture representation varies significantly across different climate models, variables like precipitation and temperature are more consistently simulated. Utilizing SPI/SPEI ensures better comparability and consistency when applying our framework across a multi-model ensemble.

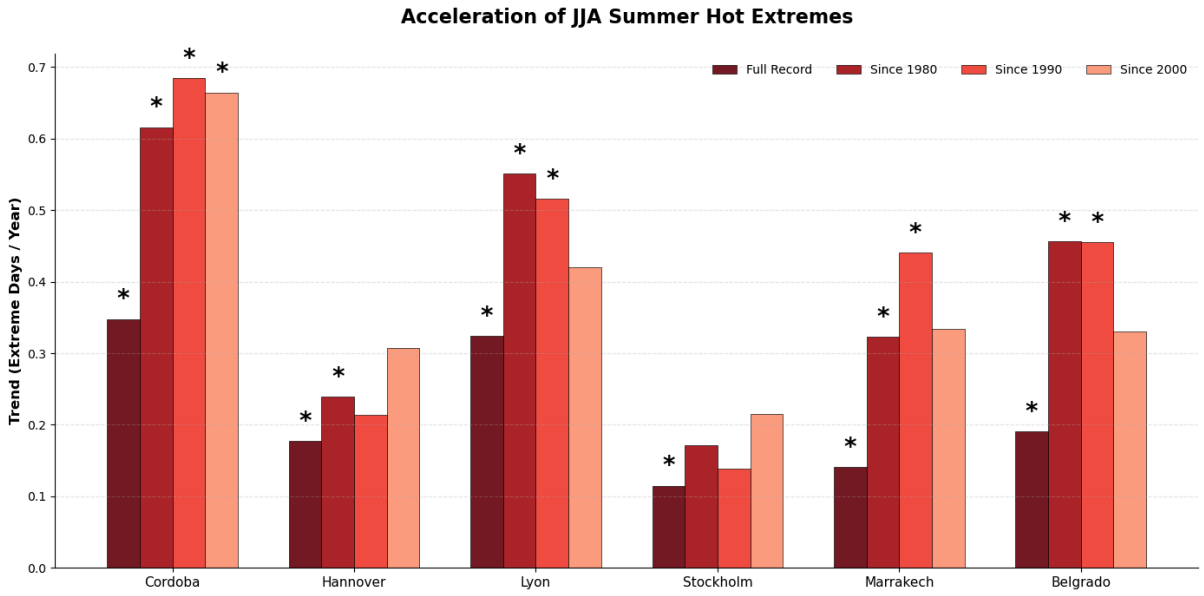
L. 165: The statement that "the number of extreme days... remains low" is confusing given that ~10% of days are extreme by construction. The intended meaning (class imbalance challenges for ML) should be clarified.

Thank you for pointing this out. We agree that the statement might be misleading. We suggest re-writting the paragraph:

The model is trained on data from 1950 to 2013. Validation years are chosen by randomly selecting contiguous blocks of years, while the testing period spans 2014 to 2024. Extreme days are defined using the 90th percentile threshold (i.e., approximately 10% of all days), which results in a class imbalance between extreme and non-extreme days (see Fig. \ref{fig: mahlstein plot cordoba}). This imbalance represents a central challenge during model training, as machine learning algorithms may otherwise default to predicting the majority (non-extreme) class if no specific strategy is used to address it. At the same time, the relatively limited number of extreme cases increases the risk of overfitting in this binary classification problem when using models with many parameters. This constraint needs to be considered when designing the model architecture and training procedure. The imbalance issue becomes even more pronounced when higher percentile thresholds are used to define extremes, as the number of available training samples decreases further.

L. 223: The claim that Belgrade shows "a visible trend after approximately 1960" likely applies to other locations as well. It would be useful to conduct formal trend analysis with significance testing for different periods (e.g., full record vs. 1980 onwards), particularly since continental locations may have experienced aerosol-related dimming effects in earlier decades.

Thank you for the suggestion. We have conducted the proposed trend analysis in the next figure:



We show trends (linear trend) computed for different periods for the different locations. Those marked with an asterisk mean that the trend is significant at the 95% confidence level. The significance of the the trend was assessed using a t-test on the slope coefficient from a linear regression. We do see an increase in trend for all locations if we compare the full period vs the period 1980 onwards. Additionally, the trend magnitude decreases for more recent periods in Cordoba, Lyon, Marrakech and Belgrado, though the trend is only significant for Córdoba.

We will include this figure in the appendix and briefly mention the analysis done in the main text:

Clear warming trends are apparent for Córdoba and Lyon regions. Belgrade region also shows a visible trend after approximately 1960. To quantify these observations, we performed a formal linear trend analysis across different time windows (see Fig. B1).

The caption of the included figure reads:

Trends in JJA Summer Hot Extremes across different time periods. Bars represent the slope coefficient (trend magnitude) of a linear regression applied to the yearly count of extreme days for four periods: the full available record (dark red), since 1980 (red), since 1990 (light red), and since 2000 (lightest red). Asterisks (*) indicate trends that are statistically significant at the 95% confidence level, determined using a t-test on the slope coefficient. A generalized acceleration in the frequency of extremes is visible when comparing the full record against the period starting in 1980. In the most recent period (since 2000), a slight decrease in the trend magnitude—and a corresponding loss of statistical significance—is observed in Lyon, Marrakech, and Belgrade. A similar slight decrease in magnitude is noted for Córdoba when compared to the 1990–onwards period, though the trend there remains statistically significant.

L. 229: "table 1 in ??"

This has been corrected.

L. 77: "sect. ??"

This has been corrected.

L. 336, 354: There appears to be duplicated/misplaced text here (the abstract seems to be repeated). Please check.

This has been corrected.

Fig. 5: See major comment above regarding CO2 heterogeneity.

We have answered the respective comment in the Main Comments section.

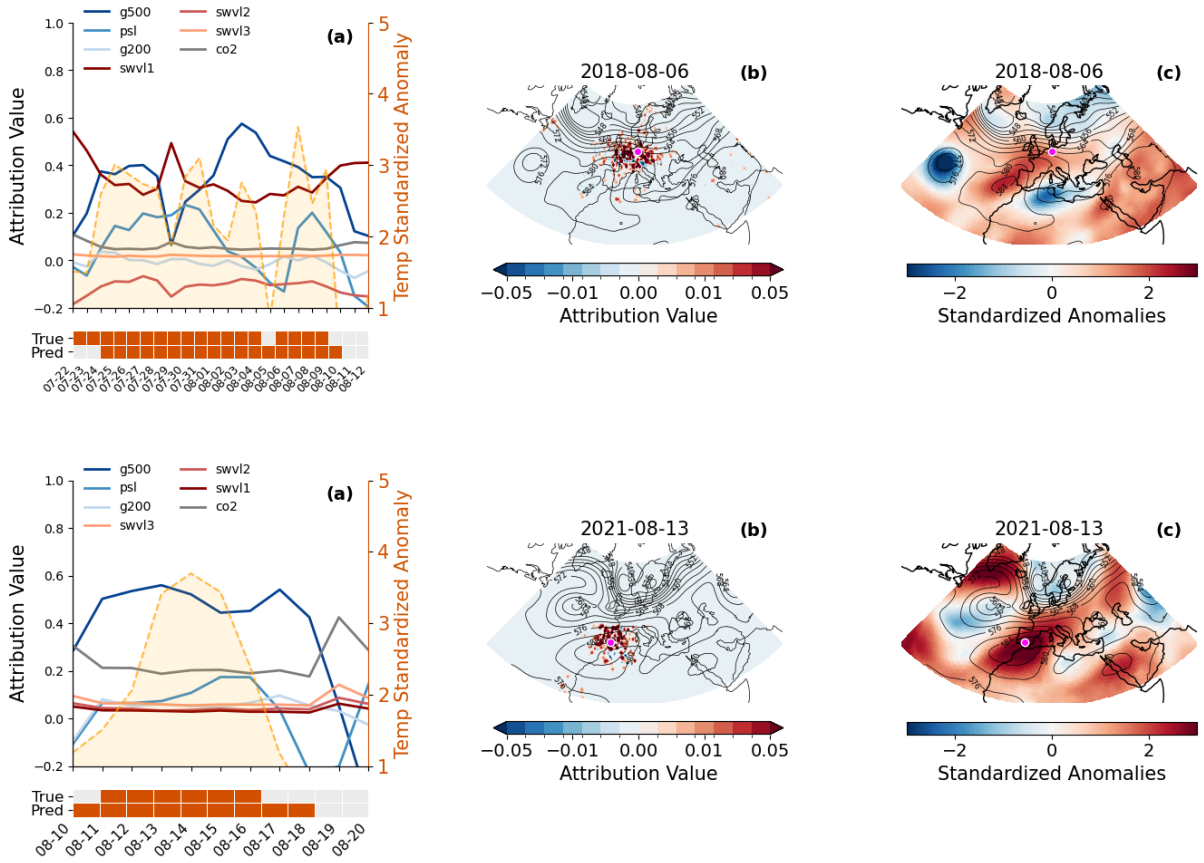
Fig. 6a: It is encouraging to see that deeper soil layers contribute more in Córdoba, consistent with top soils being fully desiccated in summer. This physical interpretation could be mentioned explicitly.

We agree on the relevance of this result. Thanks for pointing it out. We have added the following text in L277:

The importance of swvl3 in Córdoba is slightly larger than swv1 and swvl2, consistent with top soil layers being fully desiccated in summer.

Figs. 7 & 8: Showing temperature anomalies (e.g., as background shading) would help readers understand when the temperature peak occurs relative to the SHAP evolution—particularly for the Hannover case where g500 SHAP keeps increasing rather than weakening after what one might assume is the peak. Also: "standarized" is a typo in panel c, and the colour schemes differ between figures without obvious reason.

Thank you for the suggestion. We have now added the temperature anomalies as background shading in both figures. Additionally, we have increased the time period of both case studies to have a better picture of the evolution of the SHAP values for the different features. Lastly, we have also added information on the predicted and true labels (binary classification) at the bottom of panel (a). See below the final version of both figures:



We have also included suggestions made by Referee 1. The dot indicating the target location has been changed to magenta to make it more easily visible. Additionally, g500 contours have been added in panel (b).

L. 383: Given the pattern of increasing land component importance from the 80th to 90th percentile, it might be worth acknowledging that land variables could play an even more important role for more extreme events (even if the 95th percentile analysis is not robust).

Thank you for pointing this out. We have added this text in L299 with the proposed suggestion:

Land variables could play an even more important role for more restrictive percentiles like the 95th percentile. However, as mentioned in section 3.1, we do not show the explainability results for this more extreme case due to the lack of stability in the accuracy metrics when using this higher percentile.

L. 434: The finding regarding the first soil moisture level being most important should be qualified as applying to humid/transitional regions, since it clearly does not hold for drier locations like Córdoba (cf. Fig. 6a).

Thank you for the suggestion. We change the text in L434 to: *We further quantify the influence of soil moisture at three different depth levels and find that, on weather time scales, the level closest to the surface plays the largest role for humid/transitional regions, whereas the bottom layer is more important for a dry/transitional location like Córdoba.*