We answer here to the comments made by Referee 1.

We divide the comments as done by Referee 1 into Technical corrections and Specific Comments sections, with the respective subsections and comments.

Technical corrections _____

- **L77: missing section number**

The missing section number has now been added in line 77 of the revised manuscript.

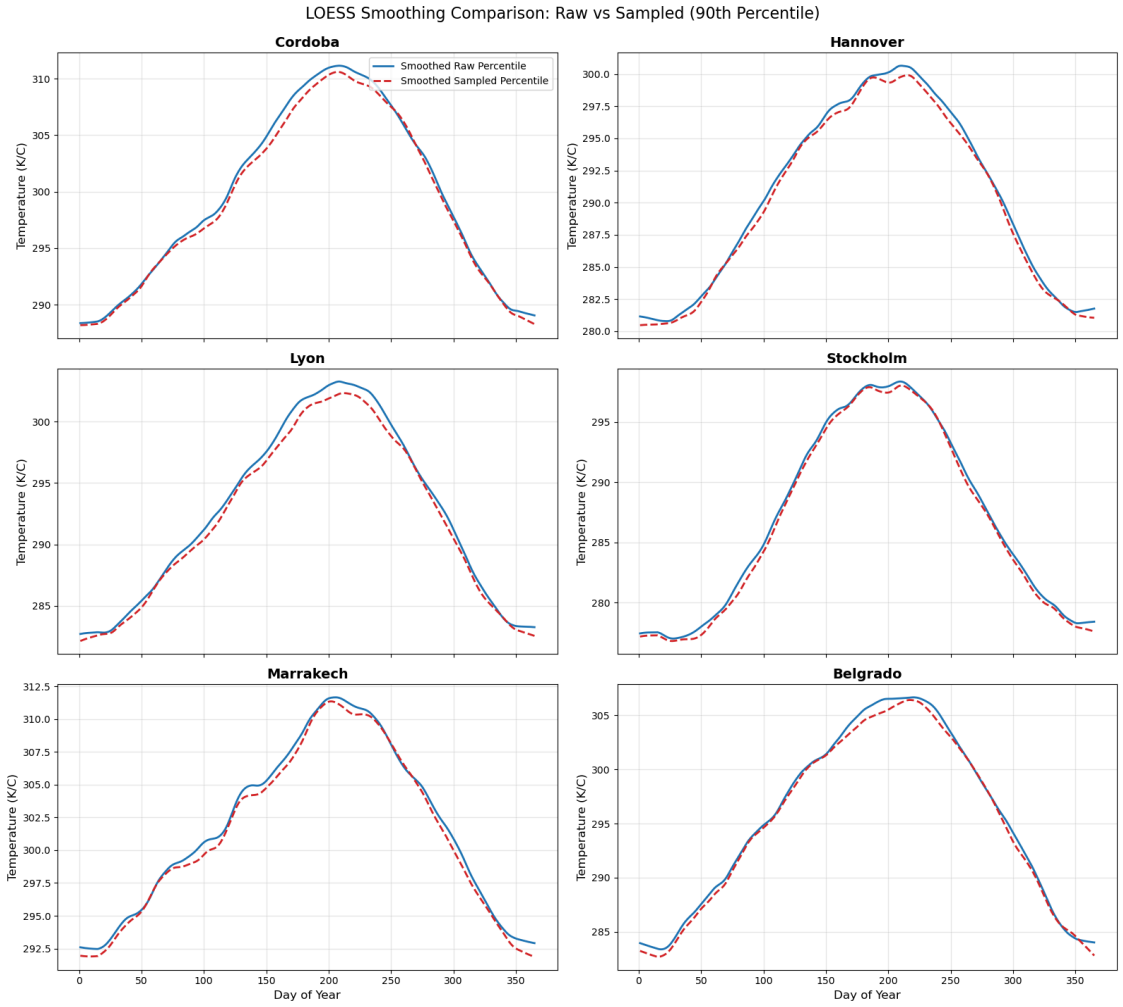
- **L85: Geopotential at 200hPa (g500) (g200)**

The name mismatch in the geopotential height variable at 200hPa has now been corrected in line 85 of the manuscript.

- **L160-161: Could the authors clarify why they use a 5 day window for the percentile calculation before a 30 day smoothing? Why not only apply the LOESS smoothing?**

Thank you for raising this point. We use a 5-day rolling window prior to the LOESS smoothing to increase the number of samples available for estimating the 90th percentile for each calendar day. Estimation of high percentiles is sensitive to sample size, and using single calendar days would provide only ~74 samples (1950–2024), which can lead to noisy and unstable percentile estimates. By constructing a centered 5-day window for each day of the year, the effective sample size is increased by a factor of five, resulting in more robust percentile estimates.

The subsequent LOESS smoothing is applied to obtain a smooth seasonal cycle from these percentile estimates. To illustrate the effect of the 5-day sampling window, we include below a figure comparing the percentile curve computed with and without the 5-day window sampling.



- **Figure 3 or in text: Please include the dimension sizes of inputs into each layers**

Thank you for the suggestion. The dimension sizes of the inputs to the different layers have now been specified in the main text (Section 2.5, line 183 in the revised manuscript).

- **Figure 3: this type of architecture design has been used in a few other studies that would be good to cite**

Thank you for the suggestion. We have added these references to the manuscript and now note that similar hybrid model architectures have been used in previous studies (Gordon et al., 2023; Mayer et al., 2024). This has been included in Section 2.5 (L193 in the revised manuscript).

- **L207: Please cite GradientExplainer**

Thank you for the suggestion. We have now added the appropriate citation for GradientExplainer (Lundberg and Lee, 2017) in Section 2.6 (line 216 in the revised manuscript). The appropriate citation to GradientExplainer is specified here: https://github.com/shap/shap?tab=readme-ov-file#citations

- **L207: I don't think the testing period has been explicitly defined yet. Please add.**

Thank you for pointing this out. We have now explicitly defined the testing period at the beginning of Section 2.5 in the revised manuscript. The testing period spans 2014 to 2024.

Text change:

We compute SHAP values in the testing period (years 2014-2024) using GradientExplainer (Lundberg and Lee, 2017b), which is designed for differentiable models, such as deep neural networks, and estimates them by using an extension of the integrated gradients method (Sundararajan et al., 2017)

- **Figure 4: This figure would be easier to read as a line plot, with a dot for each year (yellow for E-OBS and grey for ERA5-Land). This way the yellow and grey dots for each year line up vertically.**

We thank the reviewer for this suggestion. While we considered a line plot, we have chosen to retain the grouped bar plot format.

- **L229: "??" in the sentence.**

Thank you for pointing this out. The typographical error has now been corrected in the revised manuscript.

- **Figure 7b/8b:**
  - **Please include how you use/show the bootstrapping cutoffs for these figures in the caption.**

To answer this comment, we refer to the response to the second comment regarding "SHAP Results" done in the "Specific comments".

We have corrected the caption to remove the reference to bootstrapping and to explicitly describe the ensemble sign-consistency filtering and magnitude threshold applied in panel (b):

"SHAP values extreme class prediction for the Córdoba case study. a): SHAP values temporal evolution for the circulation features and the three soil moisture levels. b): SHAP spatial distribution for the

SHAP peak-day of g500 lag 1. Spatial SHAP values are shown only where at least 60% of the ensemble members agree on the sign of the attribution, indicating robust contributions across the model ensemble. Additionally, SHAP values with absolute magnitude smaller than 0.008 are masked to improve visual clarity. c): Standardized anomalies of geopotential height at 500hPa in colour and data in contours for the geopotential height at 500hPa for the predicted day (13 August 2021). The magenta dot in panel b marks the location of Córdoba."

- **Please use a different color than black for the dot locating Cordoba/Hannover. It is very hard to see.**

Thank you for the suggestion. We have changed the color of the dots locating Córdoba and Hannover to magenta to improve their visibility in the figure.

- **Figure 7c/8c: Why are the authors differentiating between geopotential and geopotential height when the difference is only a constant? I recommend only plotting g500 (which is an input into your model) as a contour and then the SHAP values could then be added as shading in this figure.**

We apologize for the inconsistent terminology. The contours shown are for the raw g500 data, not for geopotential height. Following the reviewer's suggestion, we have added the g500 contours to panel (b) to provide physical context for the SHAP values. Regarding panel (c), we have retained the standardized anomalies in the shading, as this provides essential information on the magnitude and rarity of the geopotential deviations for the selected day, while the overlaid contours now clarify the absolute state of the atmospheric field.

- **L325: Is there a citation the authors could include for the statement "contrasting soil-moisture regime"?**

Thank you for the suggestion. We have reformulated it to clarify the statement and added a supporting citation. The revised text now reads:

"We select the Hannover region for the second case study to represent a different land–atmosphere coupling regime, characterized by higher soil water content, compared to the Córdoba region (Seneviratne et al., 2006)."

Figure 1 of Seneviratne et al., 2006 illustrates the regional effect of including land–atmosphere coupling in Europe using regional climate simulations for both historical (1960–1989) and future (2070–2099) conditions.

- **L336: I think a new section should start here, as the authors are no longer specifically talking about the second case study.**

Thank you for the suggestion. We have removed mistakenly repeated text and started a new section here, as the discussion now moves beyond the second case study. This also addresses a similar comment by referee 2.

- **L354: The last paragraph repeats again here.**

Thank you for pointing this out. The repeated paragraph has now been removed in the revised manuscript.

- **Figure B1: Could you include the sample size and random chance values in this plot?**

The sample size of the extreme and non-extreme classes can be found in table B1. Regarding the random chance values, we consider this to be the naive classifier accuracy, which can be computed using the sample sizes in table B1. We do not include this accuracy in figure B1 because in this figure we want to look at the actual accruacies of the model.

Specific comments —----------

- **Training-Validation-Testing Split:**

L186-187: Could the authors clarify the 80-20 random split for the training and validation? Do the authors split the data into time chunks or are the samples randomly selected? If samples (rather than chunks) are randomly selected, how do the authors account for temporal autocorrelation in the 80-20 random split? For example, if the samples are randomly grabbed, it is possible to have 21 January 2001 in validation and 22 January 2001 in training.

The authors acknowledge that the results of the first draft were generated using a daily random train-val split, and that this implied the problems mentioned in the comment above. The train-val split method has been changed to a 80-20 random selection of years by complete year chunks. We include below the list of years used for both training and validation:

Training years: 1951, 1953, 1954, 1958–1986, 1988–1991, 1993, 1995–1998, 2000, 2002–2005, 2007, 2009, 2011–2013

Validation years: 1950, 1952, 1955, 1956, 1957, 1987, 1992, 1994, 1999, 2001, 2006, 2008, 2010

This ensures that all days within a year are consistently assigned to either training or validation, avoiding the temporal overlap issue. The explainability results were not notably affected by this change, but we do have noticed an improvement in the stability of the results across different seed runs in terms of accuracy.

Table B1 now includes the list of the training and validation years at the bottom.

L186-191: Is the training-validation 80-20 random split from the 1950-2014 training period? If so, I recommend changing Table B1 to also have the random 20% that

represents the validation data. I also recommend clarifying in the text that 1950-2014 includes training *and* validation data.

Yes, the validation samples are selected from the 1950–2013 period. Table B1 has been updated to include the counts of extreme and non-extreme classes for the randomly selected validation years. In addition, the text has been clarified to specify the years used for the train–validation split. The revised sentence at the beginning of Section 2.5 now reads:

"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."

- **Network confidence**
    - **L232-233: Do you expect accuracy to increase with confidence because you believe some events are actually more predictable than others or do you think this is just a factor of the loss function?**

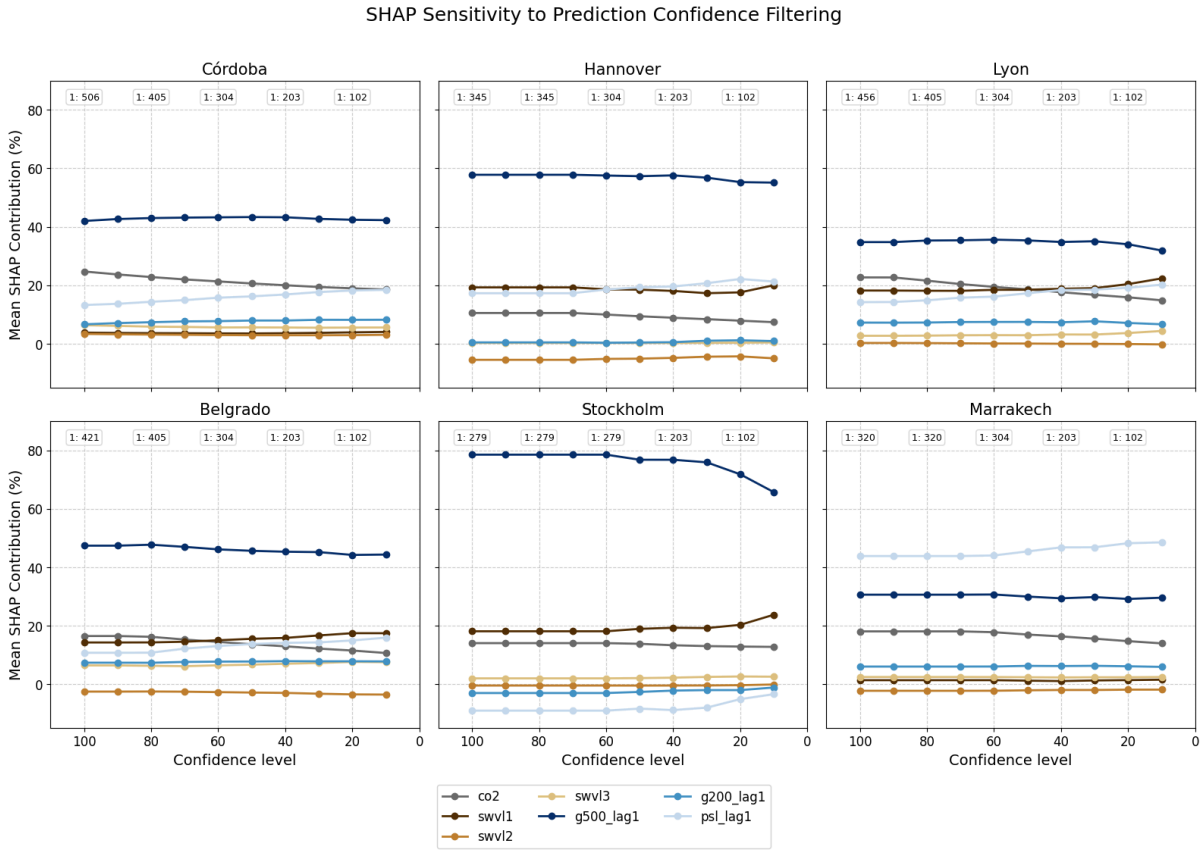
We think that the model predicts some events with higher probability because the role of the drivers is clearer in some events than others. For instance, in the Córdoba case, a strong anticyclone over the Iberian Peninsula, plus dry conditions in soil moisture locally, would be the perfect combination for triggering a hot extreme in this location. The model might easily learn that these conditions are suitable for having an extreme event. However, a variety of conditions in the different drivers might be suitable in a specific location for developing an extreme.

When we filter the results by confidence, we do so in both classes so that we have the same class imbalance as without filtering. For this reason, we do not think that the accuracy increase is a factor of the loss function.

    - **Why do the authors choose to look at the 50% most confident extreme predictions**

Thank you for pointing this out. In the figure below we can see that the mean SHAP contributions for the primary drivers remain remarkably stable across the various confidence levels. This consistency suggests that the core physical mechanisms identified by the model are robust and that the specific choice of threshold does not fundamentally alter the interpretation of the results.

However, we observe slight shifts in feature importance as confidence increases. By focusing on the 50% most confident cases, we proactively reduce potential 'noise' coming from lower-confidence predictions. We opted not to implement a more restrictive threshold (such as the top 20% or 10%) to ensure that our mean results are calculated from a sufficiently large and representative sample of events. We show in the plot the number of samples for the extreme class (label 1) for the [100,80,60,40] percentages of top confident predictions.



- **Choice of Baseline**

  L203: Could the authors explain why they chose a baseline of the average predictions? Given the authors are interested in the drivers of hot temperature extremes, I think it may be more useful to direct the XAI method to answer the question "which regions made the network predict an extreme *as opposed to a non-extreme*?" If this is the case, I think the authors should use the average non-extreme days for their baseline. Phrased another way, the current baseline choice is answering "what regions make this different from the average day?" Below are some relevant citations to explore this concept further. If the authors chose to keep the baseline as the average of all the training data, please include reasoning and citations.

  - **Mamalakis et al. (2023). Carefully Choose the Baseline: Lessons Learned from Applying XAI Attribution Methods for Regression Tasks in Geoscience. https://doi.org/10.1175/AIES-D-22-0058.1**
  - **Sundararajan et al. (2017): Axiomatic Attribution for Deep Networks. https://arxiv.org/pdf/1703.01365**

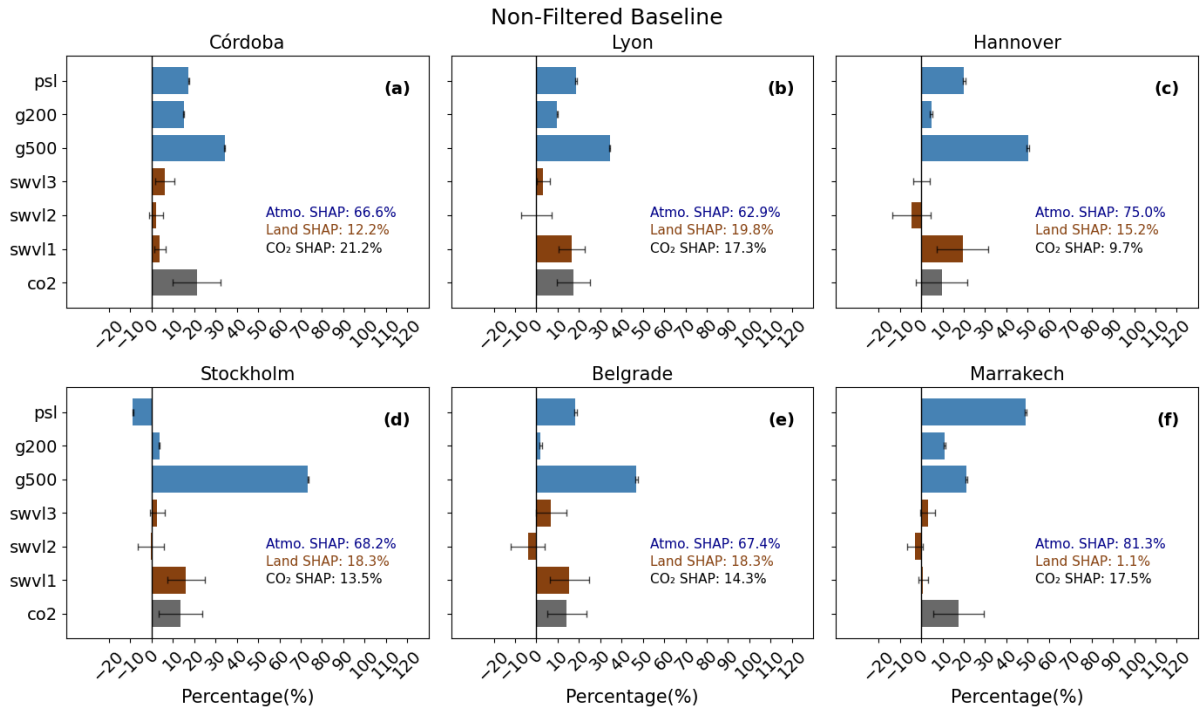
Thank you for the insightful comment. We conducted a test to assess the effect of using a baseline comprising only non-extreme samples (see plot below). In this comparison,

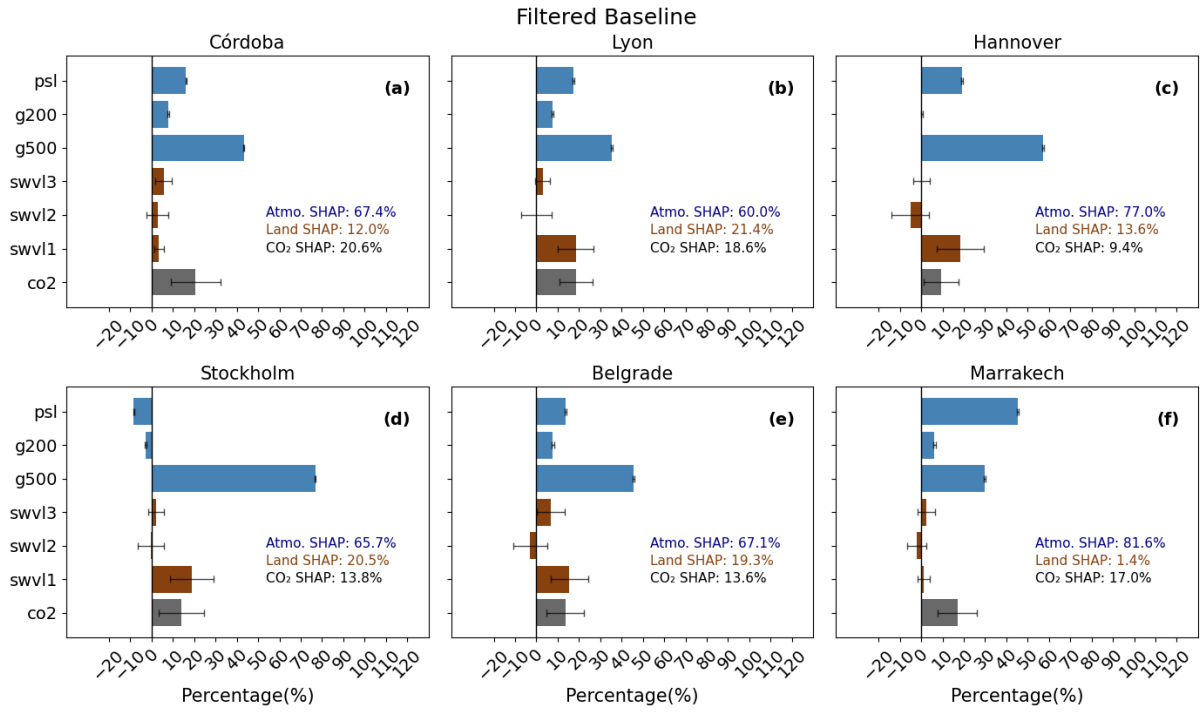
the “filtered baseline” case uses only non-extreme days, while the “non-filtered baseline” includes both extreme and non-extreme days randomly selected from the training period.

The results indicate little difference between the two approaches. Nonetheless, we agree with the rationale for using a baseline consisting solely of non-extreme days and have adopted this modification in the code. Correspondingly, the manuscript has been updated at the end of Section 2.6 (line 220) as follows:

Additionally, the method requires selecting a baseline to serve as a reference in the SHAP computation. Following best practices suggested by Mamalakis et al. (2023), the baseline was constructed by randomly selecting 300 samples from the training period that had non-extreme labels. Using only non-extreme days for the baseline is appropriate in this study, as our focus is on identifying the drivers of predicted extreme events with the aforementioned ML model.

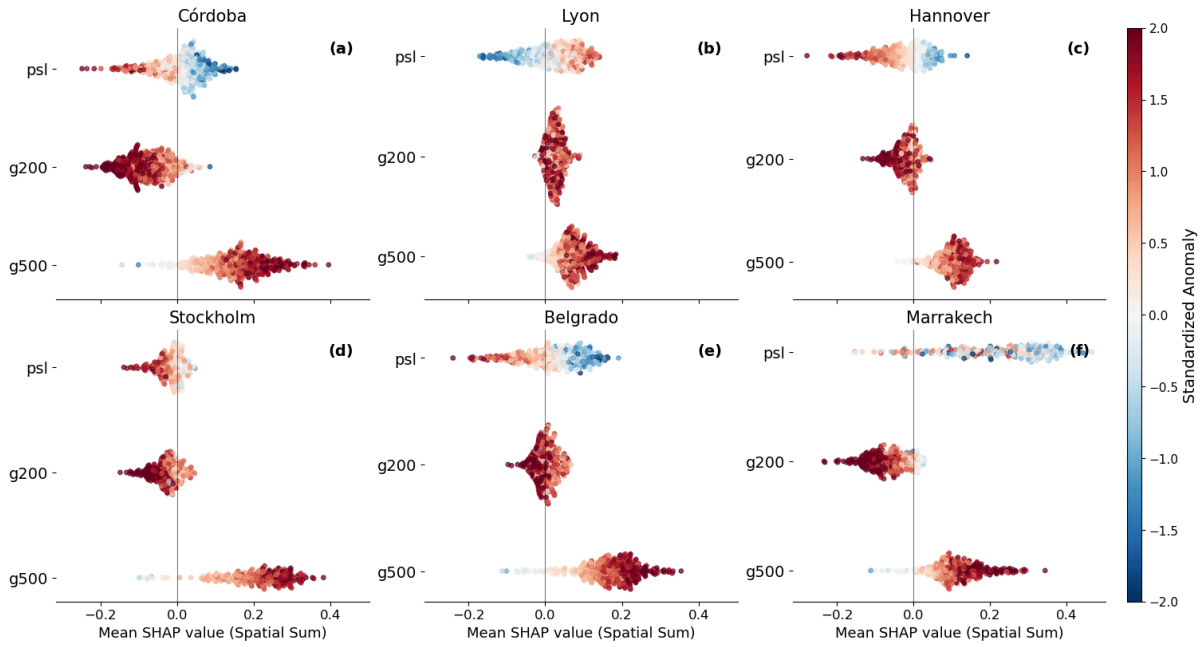
This change also includes the citation to Mamalakis et al. (2023), as suggested, to justify the baseline selection.





- **SHAP results**
  - **Figure 6: Could you also include this figure for circulation?**

Thank you for your suggestion. To create a similar figure for the circulation variables, we need to consider the spatial dimensions of these, which is not the case for the local scale variables. To create an equivalent figure for the circulation, we have summed up the SHAP values and averaged the standardized anomalies of the circulation variables in a 400km by 400km region surrounding the target location. We think this approach is reasonable considering that the whole EuroAtlantic spatial region is used for the fields. The resulting plot is the following:



- **L313: Please add more detail on how bootstrapping was done.**

Thank you for pointing this out. The reference to "bootstrapping cutoffs" in the caption was inherited from an earlier version of the manuscript and was inadvertently left in the text. No bootstrap-based filtering is applied in this figure.

To improve the visual interpretability of the map, we apply a small magnitude threshold to the SHAP values, masking values with absolute magnitude smaller than 0.008, which removes very weak contributions that are not visually meaningful.

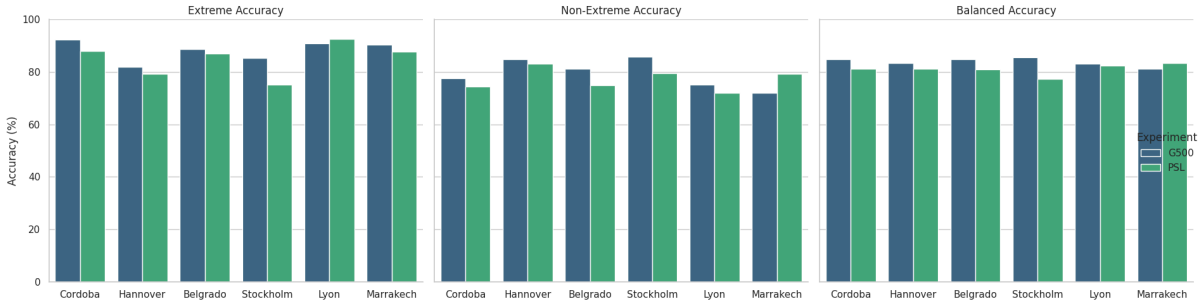
The text now reads:

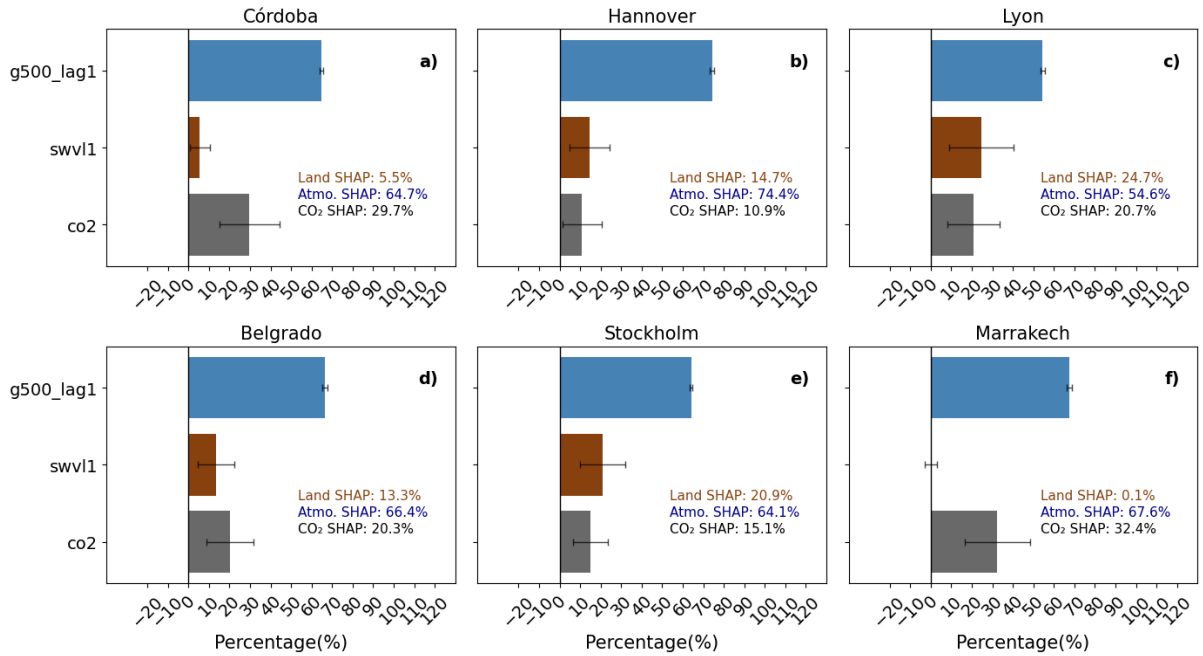
"Fig. 7b shows the SHAP values for the g500 feature at the SHAP peak identified in Fig. 7a. Grid points with |SHAP| < 0.008 are masked out for clarity."

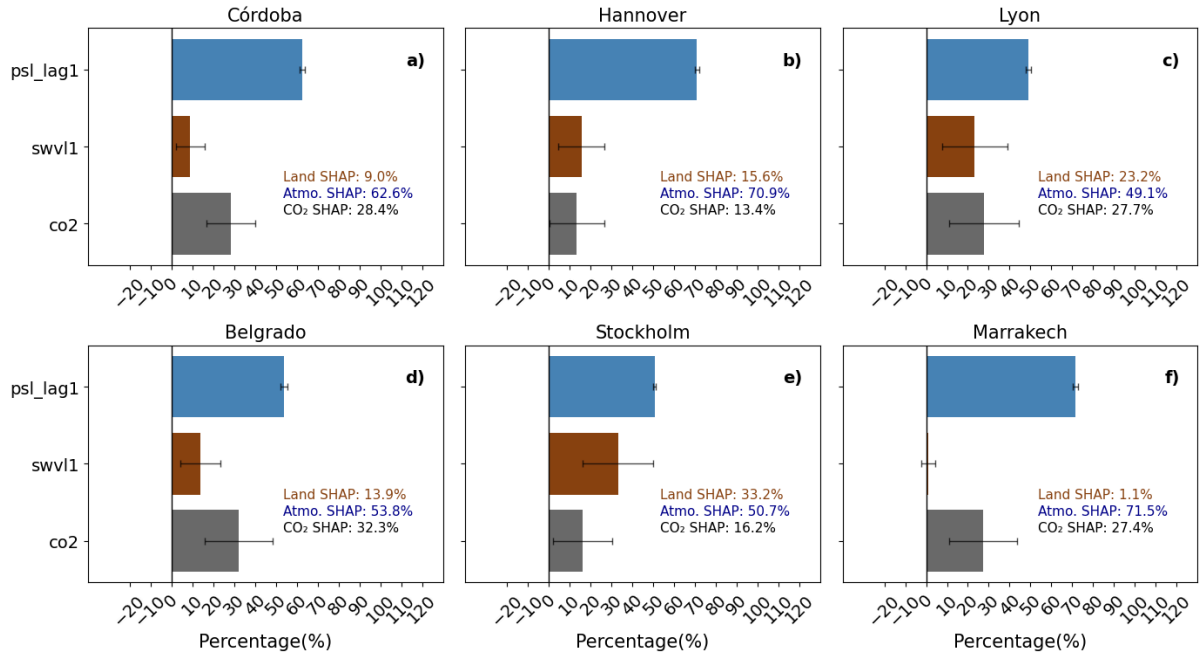
- **Could the authors discuss why they think the SHAP values for PSL and g500 don't evolve the same for the case studies? I would expect them to contribute similar information. Is PSL redundant and therefore, not used by the model? Could the same information be retrieved from PSL if g500 wasn't included?**

Thank you for your comment. To analyse this, we have decided to compare the accuracies and mean SHAP values over time for two models, one trained with the features [co2, swvl1, psl] and one with the features [co2, swvl1, g500].

The following two plots show the accuracy comparison and the mean SHAP values (Fig.5 in the main text)


Extreme Accuracy
Non-Extreme Accuracy
Balanced Accuracy
Accuracy (%)
Experiment
G500
PSL
Cordoba
Hannover
Belgrado
Stockholm
Lyon
Marrakech


Córdoba
Hannover
Lyon
Belgrado
Stockholm
Marrakech
g500_lag1
swvl1
co2
a)
b)
c)
d)
e)
f)
Land SHAP: 5.5%
Atmo. SHAP: 64.7%
CO₂ SHAP: 29.7%
Land SHAP: 14.7%
Atmo. SHAP: 74.4%
CO₂ SHAP: 10.9%
Land SHAP: 24.7%
Atmo. SHAP: 54.6%
CO₂ SHAP: 20.7%
Land SHAP: 13.3%
Atmo. SHAP: 66.4%
CO₂ SHAP: 20.3%
Land SHAP: 20.9%
Atmo. SHAP: 64.1%
CO₂ SHAP: 15.1%
Land SHAP: 0.1%
Atmo. SHAP: 67.6%
CO₂ SHAP: 32.4%
Percentage(%)


Córdoba
Hannover
Lyon
Belgrado
Stockholm
Marrakech
psl_lag1
swvl1
co2
a)
b)
c)
d)
e)
f)
Land SHAP: 9.0%
Atmo. SHAP: 62.6%
CO₂ SHAP: 28.4%
Land SHAP: 15.6%
Atmo. SHAP: 70.9%
CO₂ SHAP: 13.4%
Land SHAP: 23.2%
Atmo. SHAP: 49.1%
CO₂ SHAP: 27.7%
Land SHAP: 13.9%
Atmo. SHAP: 53.8%
CO₂ SHAP: 32.3%
Land SHAP: 33.2%
Atmo. SHAP: 50.7%
CO₂ SHAP: 16.2%
Land SHAP: 1.1%
Atmo. SHAP: 71.5%
CO₂ SHAP: 27.4%
Percentage(%)

Regarding the accuracy comparisons, we see in the figure that generally the accuracies are similar for both classes and also the balanced accuracy. Except for Marrakech, all locations show a slightly higher accuracy for the g500 case. For Marrakech, the non-extreme accuracy and the balanced accuracy are a bit higher for the psl case. This is consistent psl mean SHAP being higher for Marrakech when the three atmospheric fields are used to train the model (Fig.5 in the main text). The g500 field carries more information on the atmospheric state, so we think it is reasonable that the accuracy is generally higher for the g500 case.

As for the mean SHAP results, the Land contribution for Stockholm changes significantly, being twice as large in the psl case compared to the g500 case. In conclusion, While the similar accuracy levels across both experiments confirm that psl can indeed retrieve the necessary predictive information in the absence of g500, the shift in SHAP values—particularly in Stockholm—reveals how the model compensates. In the psl experiment, the model increases its reliance on land-surface information (swvl1) to maintain accuracy. This suggests that g500 better captures the atmospheric drivers of extremes directly, whereas PSL provides a noisier signal that requires the model to lean more heavily on land-state persistence to achieve similar performance. We think that it is important to note that in the case of including all three atmospheric variables, in some cases (like in the Marrakech model), the model switches the importance from g500 to psl, suggesting in some cases psl provides a stronger signal to predict the extreme events.

- **In a similar line of thinking, do the authors think temperature extremes could be predicted using *only* soil moisture information?**

Thank you for the question. We trained the MLP branch of the model using only CO_2 and soil moisture inputs to predict temperature extremes. In this case, the model's accuracy did not exceed 70% for either class. These results suggest that soil moisture alone is insufficient to predict hot temperature extremes at a daily timescale, and that suitable atmospheric conditions are also required. The XAI analysis reflects the modulating effect of soil moisture on extreme events, rather than being a standalone predictor. We think soil moisture could still be a suitable proxy to predict weekly/monthly counts of extreme events.

- **Do the authors think temperature extremes could be predicted using *only* CO2 information?**

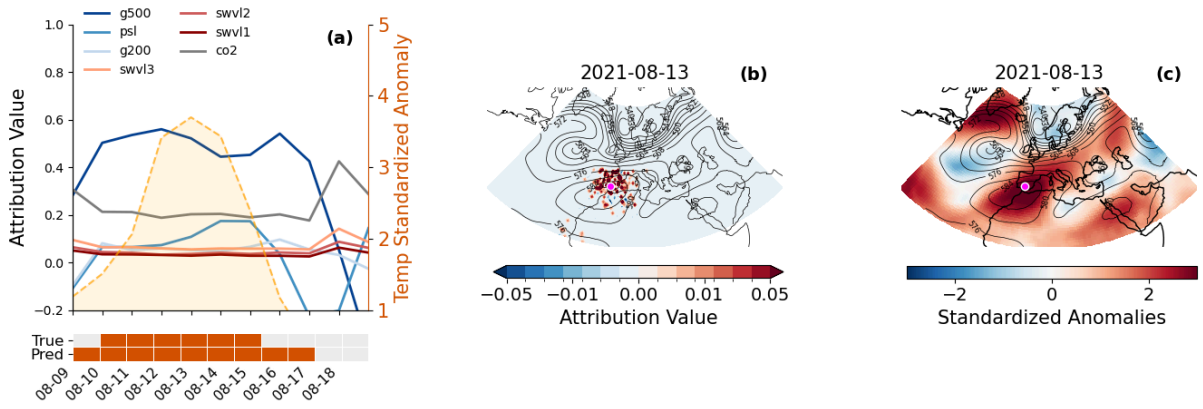
Following the reasoning in the previous comment regarding soil moisture, we also tested whether CO_2 alone could predict daily temperature extremes. We trained a simple MLP using only seasonally averaged CO_2 data to predict the binary classification. Due to severe class imbalance, the model either predicted mostly the majority class or, when class weights were applied, defaulted to the extreme class. These results indicate that CO_2 alone is insufficient to predict daily hot extremes. We think this result is expected,

since a seasonal mean CO2 value cannot predict the daily variability of hot extremes classification. However, CO_2 may still be useful for predicting yearly mean counts of hot extremes, which exhibit a visible trend (Fig. 4).

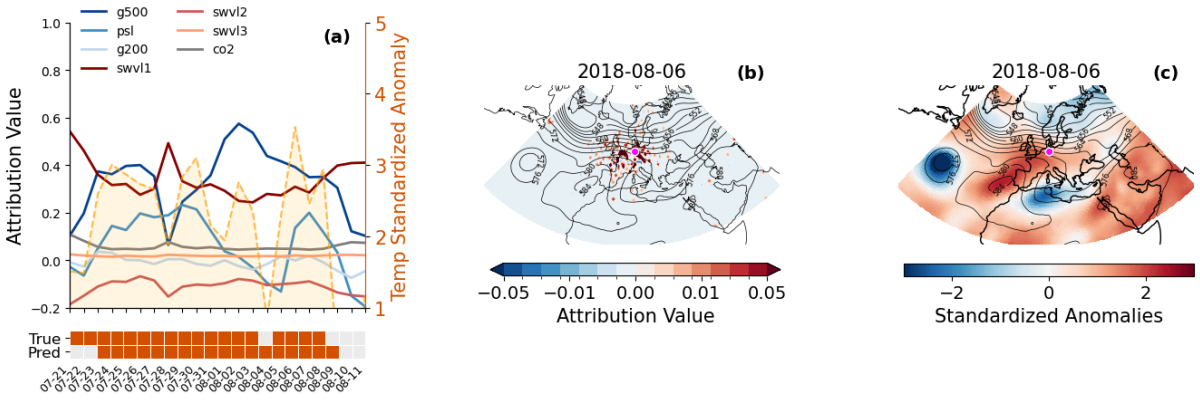
- **Figure 7b-c/8b-c: Are the authors looking at the circulation predictor of the extreme heat (e.g. one day before the peak day) or the circulation on the peak day? Given the set up of the model, the authors should look at the *day before* SHAP values as this is what is associated with the prediction of the peak day. Otherwise, these plots are highlighting the regions important for predicting the day after the peak (14 August 2021 and 5 August 2018, respectively).**

We clarify that the SHAP values shown for the temperature peak day represent the importance of the circulation state from the previous day. Following the reviewer's suggestion, we have added the temperature anomalies as background shading in the temporal evolution plots (Figs. 7 and 8). This allows for a direct comparison between the temperature peak and the SHAP value progression. We have additionally included on the bottom of panel a the True and Predicted of the binary classification, which we believe also helps to interpret the time evolution of the SHAP values shown in panel (a). Lastly, we have extended the period of both case studies to correctly see the temporal evolution of the SHAP values at the beginning and ending of the main extreme event.

The date on panels (b) and (c) stands for the date of the temperature peak. The contours and standardized anomalies of g500 shown are those of lag1, meaning the atmospheric field conditions on the 2021-08-12, and the same applies to the SHAP values shown in panel (b).



Comments Figure: 6. Figure 7 in the main text.



- **Are these case studies part of the 50% most confident predictions? If not, why do you think that is?**

Thank you for raising this point. For both case studies, the time steps that correspond to true positives (i.e., when the model correctly predicts the event) fall within the 50% most confident predictions of the model. This indicates that when the model successfully identifies the event, it generally does so with relatively high confidence.

More specifically, in the Hannover case study (21 July–11 August 2018), the majority of the days during the core phase of the event fall within the set of the 50% most confident predictions. Only the first two days (21–22 July) and the final three days (9–11 August) fall outside this subset. Similarly, in the Córdoba case study (9–19 August 2021), most of the central days of the event are within the most confident predictions, while the first day and the final three days fall outside this subset. This suggests that higher confidence is seen on peak events days while lower is observed on onsets or fading phases.

This behavior suggests that the model tends to be most confident during the core phase of the heatwave events. In contrast, the model appears to struggle more with the transition periods at the beginning and end of the events, where the signal may be weaker or more ambiguous. This is reflected in the lower confidence and occasional misclassifications during those periods in both case studies.

- **L427-228: “SHAP values, however, remain relatively unexplored for such pattern identification” - this statement is not true. There are many papers which use SHAP values to explore regional importance. A few examples are below:**

We thank the reviewer for pointing out these relevant studies and for providing the additional references. We agree that the initial statement was too broad. We have re-written that sentence and included two of the suggested references instead in L427.

We do not include the Van Straaten et al. 2023 work because in that study they do not spatial patterns of their computed SHAP values.

The text now reads:

While recent studies have increasingly employed XAI to identify regional climate drivers and sources of predictability (e.g., Zhang et al. (2024); Mamalakis et al. (2023b)), our work specifically leverages XAI to disentangle the relative roles of atmospheric circulation and land-surface conditions in driving localized extreme events.

- **Statements in L395 and L430 appear counterintuitive. Could the authors clarify the discrepancy in which they find that in the arid regions (e.g. Cordoba and Marrakech) the land importance was negligible, but previous research has shown "that dry soils intensify and prolong heat extremes"? The statement in L430 is confirmed by the anti-correlated SHAP values and soil moisture values, but is contradictory to the northward gradient of soil-moisture importance. Is the statement in L430 only true in specific climates? Please clarify.**

The authors would like to clarify the complementarity of the two statements. The attribution study indicates that, overall, the land component contributes positively to the prediction of heat extremes, with varying magnitudes across locations and a noticeable northward gradient when comparing Marrakech and Córdoba to more northern sites. For Cordoba, the deepest layer has a more important role, consistent with upper layers being dried out in summer and therefore showing no variability, as noted by Referee 2. As noted in the review comment, Figure 6 shows the anti-correlated SHAP and soil moisture values, but also reveals differences in the spread of SHAP magnitudes across locations, reflecting the aforementioned northward gradient—particularly for the soil moisture layer closest to the surface. Additionally, in light of the changed values of Land importance shown in Figure 5 after incorporating some of the reviewers' recommended changes, the authors consider that the overall Land SHAP for Córdoba is not negligible, even though it is smaller than that of more northern locations."