Response to Anonymous Referee #2 on Manuscript ID EGUSPHERE-2026-832

GENERAL COMMENTS

This manuscript presents a timely and interesting contribution to the statistical modeling of heat-related extremes in Spain. The authors develop a hierarchical spatio-temporal logistic-regression framework to model the occurrence of daily maximum-temperature calendar-day records using ERA5 geopotential covariates at 300, 500, and 700 hPa. The topic is relevant to NHESS because record-breaking temperatures are high-impact climate hazards with implications for public health, agriculture, drought stress, wildfire risk, energy demand, and heat-stress management.

The strengths of the manuscript are the use of a long observational record over Spain, the focus on calendar-day records with well-known theoretical properties under stationarity, the interpretable modeling strategy, and the attempt to link surface temperature records to large-scale atmospheric circulation. The open repository is also a positive aspect for reproducibility.

However, several methodological and presentation issues should be addressed before publication. In particular, the manuscript needs a clearer treatment of temporal non-stationarity, the intrinsic time dependence of record probabilities, calibration of rare-event probabilities, uncertainty in variable selection, and consistency in the training/testing periods. I therefore recommend major revisions.

Response: We sincerely thank the referee for the positive and constructive assessment of our manuscript. We are pleased that the referee found our work to be timely, interesting, and highly relevant to the scope of NHESS. We also appreciate the recognition of the strengths of our study. We take the referee's methodological and presentation concerns seriously, and we agree that addressing these points will significantly strengthen the rigor and clarity of our framework.

SPECIFIC COMMENTS

1. The manuscript is relevant to natural hazards, but the introduction and discussion should make the hazard link clearer. Record-breaking daily maximum temperatures should be framed not only as a statistical climate indicator, but also as an operationally relevant heat-hazard signal. The authors should explain how this record-based approach complements classical heatwave definitions based on absolute or percentile thresholds.

Response: We agree that stronger framing within the natural hazard context is necessary. We will include an explanation in the Introduction and Discussion sections of how this record-based approach complements classical heatwave definitions based on absolute or percentile thresholds.

In the Introduction (lines 24–25), we mention that many meteorological services use record-breaking events as climate indices, as they provide a valuable alternative approach for analyzing extreme behavior. Our intention was to indicate that meteorological agencies already use record-breaking events as one indicator of heat hazard. Nevertheless, we agree that this connection is not stated explicitly enough, and we will add additional remarks in the Introduction to make this point clearer.

We will reconstruct the paragraph from the introduction located between lines 63 and 68, strengthening the idea of the link between heat-hazards and our method. It will read as follows: "Understanding and quantifying the influence of geopotential covariates on the occurrence of record-breaking temperatures is essential for conducting climate attribution studies. ***In this context, one of the objectives of this article is to propose a tool to study the deviation of record occurrences from the stationary baseline. The proposed methodology analyzes and explains the behavior of natural hazards by linking records to the atmospheric covariates that drive them.*** *The methods would also be useful to develop statistical downscaling methods to project the future frequency of record-breaking temperatures under global warming scenarios. One might argue that such projections could be derived from daily temperature projections, but these often fail to accurately capture the behavior of temperature extremes and, in particular, record events, the most extreme observations. Therefore, generating reliable projections for record-breaking temperatures requires specific downscaling methods (Castillo-Mateo et al., 2025a)."*

In the paragraph that starts on line 34 and ends on line 36 we will add an explanation to emphasize how our approximation based on records complements other techniques based on quantile thresholds. At the end of this paragraph, this text would be included: ***"The proposed methodology complements other options, like analyzing changes in extremes based on extreme heat events defined by a threshold, which can be set empirically or might require defining the distribution of a variable. There are two advantages to our approach: first, the researcher does not need to choose or define a threshold, and second, the analysis of records is non-parametric."***

We will clarify the link with heat hazards in the Discussion section by adding the following sentence to the paragraph starting at line 447: "Our study represents the first

attempt to characterize T_x records across peninsular Spain using geopotential at several pressure levels as predictors. ***We have shown that record-breaking events can be viewed as another expression of heat hazard. This provides a complementary perspective that may be useful for the analyses carried out by meteorological agencies.*** *Our results demonstrate a consistent statistical–physical link between upper-level atmospheric circulation (…)"*

2. The title and abstract refer to 1960–2023, but the data description states that the testing period covers 2012–2024, while later sections refer to validation over 2011–2023. This is confusing and must be corrected. Please clearly state the exact observational period, training years, validation/testing years, and whether 2024 is included or not.

   Response: Thank you for pointing out this inconsistency; you are entirely right. The mention of the 2012–2024 period was an error, and the year 2024 is not included in the study. The total observational period spans 1960–2023, with 1960–2010 used for training and 2011–2023 for testing. In the revised manuscript, we will correct this text in lines 104 and 105 so it reads: "The training period spans 1960 to ***2010*** *(approximately 80% of the data), while the testing period covers* ***2011*** *to* ***2023****."*

3. Include a baseline record-probability structure: a key property of record events is that, under stationarity and independence, the probability of a record at time t is $1/t$. The manuscript discusses this in the exploratory analysis, but it is not clear whether the final logistic models include this baseline record-age effect. Please compare the proposed model against at least three baselines: (i) a stationary $1/t$ record-probability model; (ii) a time-trend-only or year-smooth model; and (iii) a geopotential-only model. This would show the added value of the atmospheric covariates beyond the changing record baseline and the long-term warming trend.

   This suggestion is very useful for us as it allows us to show the variability explained by our modeling. We want to clarify that the final logistic models do not include the baseline record-age effect.

   We have adjusted (i) a model with an offset of $-\log(t-1)$, which corresponds to the stationary model. This is because under a stationary climate, we assume the probability of a record occurring at year t to be $p_t = 1/t$. In a logistic model, the logit link transforms a probability (p) into a log-odds value using the following formula:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Plugging the stationary record probability ($p_t = 1/t$) into this function, we get:

$$\begin{aligned}
\text{logit}(p_t) &= \log\left(\frac{1/t}{1-1/t}\right) \\
&= \log\left(\frac{1/t}{(t-1)/t}\right) \\
&= \log\left(\frac{1}{t-1}\right) \\
&= -\log(t-1)
\end{aligned}$$

(ii) A model that captures the temporal trend by including the term $\log(t-1)$ *freely as a predictor rather than a constrained offset. (iii) Our selected model M2, which relies exclusively on geopotential variables but additionally includes interactions with latitude and longitude (which are time-constants). Anticipating the next comment, we also consider the M2 model using detrended variables. We do not include the version based on standardized anomalies, since standardization rescales the predictors by a constant factor and therefore yields an equivalent model.*

Table R1 summarizes the results for the models considered: the stationary model (Stationary), the trend-only model (Trend), our selected model (M2), and the corresponding model fitted using detrended variables (M2.detrend). For each model, we report the number of parameters (k), the number of variables (n_{var}), the AIC, AUC and the precision–recall AUC (PR_{AUC}).

The stationary and trend-only models both achieve an AUC of 0.58, indicating that neither of them have sufficient discriminatory power. In contrast, the selected model (M2) achieves an AUC of 0.89. M2 also substantially reduces the AIC compared to both the stationary and trend-only models, indicating a considerably better fit to the data. This demonstrates that the large-scale atmospheric circulation contains more information on the occurrence of record events than the temporal trend alone.

When the geopotential predictors are detrended, the resulting model exhibits a loss of performance, as reflected by the increase in AIC and the reduction in PR_{AUC}. However, its discriminatory ability remains the same as M2 (AUC = 0.89). This suggests the selected predictors retain predictive information after detrending. Therefore, the relationship identified by M2 cannot be only attributed to the long-term warming trends, but also reflects the contribution of the large-scale atmospheric circulation to the occurrence of records.

Table R1: Results for the alternative models.

Model	k	n_{var}	AIC	AUC	PR_{AUC}
Stationary	0	0	82157	0.58	0.03
Trend	2	1	81693	0.58	0.03
M2	41	20	76811	0.89	0.31
M2.detrend	41	20	77710	0.89	0.29

We will include Table R1 in the supplementary material, and a new paragraph at the end of subsection 5.1.3 that will read as follows:

To evaluate the added value of our modeling approach, we compared the predictive performance of the selected M2 model against a stationary baseline model and a temporal trend-only model (Table S? in Supplementary material). Both baseline models exhibited limited discriminatory power, yielding an AUC of 0.58. In addition, M2 achieved a substantially lower AIC than both models, indicating a better fit to the data. We also evaluated M2 using detrended geopotential variables to ensure the model is not capturing long-term warming rather than a dynamically meaningful circulation-record relationship. While detrending caused an increase in AIC and a decreasae in precision-recall AUC, the model retained the discriminatory power (AUC = 0.89). Therefore, these results suggest that the relationship identified by M2 cannot be only attributed to the long-term warming trend. Although detrending results in a decrease in overall model performance, indicating that the long-term warming trend contributes to the model fit, the retained predictive performance reflects the contribution of large-scale atmospheric circulation to the occurrence of record events.

4. The manuscript shows that both temperature records and geopotentials exhibit upward trends. This creates a possible confounding issue: the model may partly capture long-term warming rather than a dynamically meaningful circulation–record relationship. Please test whether the selected geopotential predictors retain skill after detrending or anomaly-standardizing them relative to a moving climatology. A useful sensitivity analysis would compare raw geopotentials, standardized anomalies, detrended geopotentials, and models including an explicit year effect.

   Reviewer #2 is right in pointing out that the yearly aggregated evolution of T_x and geopotentials is closely related, as suggested by Figure S.1 in the Supplementary Material. Nevertheless, their relationship is more complex, as discussed in the previous

comment, where we compare models based on raw and detrended geopotential predictors, together with stationary and trend-only models. As discussed there, the detrended model retains most of its predictive skill, indicating that the relationship between geopotential and record events is not solely explained by their common long-term trends.

5. The manuscript relies strongly on AUC. AUC is useful but not sufficient for rare-event prediction, especially when non-records dominate the dataset. Please add calibration and rare-event metrics, including Brier score, log score, reliability diagrams, calibration slope/intercept, precision–recall AUC, and sensitivity/specificity at selected thresholds. If the model is proposed as a downscaling or prediction tool, calibration is as important as discrimination.

   Response: It is important to note that the proposed methodology for constructing the models does not use or depend on the AUC. The AUC is only employed to evaluate predictive performance during the validation period. We thank the reviewer for highlighting the importance of complementing the AUC with additional validation metrics. The AUC was included because it is a widely used measure of discrimination that considers all possible classification thresholds. Nevertheless, we agree that additional metrics might provide valuable information about model performance. Following the reviewer's suggestion, we expanded our evaluation metrics to include comprehensive rare-event and calibration metrics (Brier score, log score, reliability diagrams, calibration intercepts/slopes, precision-recall AUC, and threshold-specific performance). These metrics were evaluated over the validation period (2011–2023). These results for the models M1, M2, M3, and M4 are shown in Table R2 and Figure R1. Recall that these models were introduced in Section 4.1 of the manuscript. Key insights from these newly added metrics include:

Table R2: Summary of validation metrics.

Model	k	n_{var}	AUC	PR_{AUC}	Brier	Log	Cal_{Int}	Cal_{Slo}	Sens(0.1)	Spec(0.1)	Sens(0.2)	Spec(0.2)	AUC.coast	AUC.inner	AIC
M1	18	15	0.866	0.255	0.039	0.159	-0.957	1.607	0.876	0.697	0.629	0.892	0.780	0.918	79812
M2	41	21	0.887	0.313	0.037	0.152	-0.963	1.645	0.885	0.711	0.686	0.893	0.802	0.934	76821
M3	37	20	0.884	0.308	0.038	0.153	-0.975	1.619	0.882	0.711	0.643	0.900	0.804	0.921	77239
M4	30	18	0.877	0.307	0.038	0.155	-0.973	1.615	0.873	0.704	0.629	0.900	0.787	0.920	78446

- *PR-AUC: Because our positive class prevalence is only* 4.23%*, a random classifier yields a baseline PR-AUC of* 0.0423*. Our selected model M2 achieves a PR-AUC of 0.313, corresponding to a PR-AUC 7.4 times higher than the random baseline, also outperforming the simpler model M1 (PR-AUC = 0.255).*

- *Brier Score: The Brier scores are low for all models, indicating good overall probabilistic predictions. Differences between models are small, although M2 achieves the lowest Brier score (0.037).*
- *Log Score: Considering the positive class prevalence of 4.23%, the baseline Log Score for a naive classifier is 0.1752. This metric decreases to 0.159 for M1 and to 0.152 for M2, indicating an improvement in the probabilistic predictions compared with both the baseline classifier and M1.*
- *Reliability diagrams: The reliability diagrams for models M1, M2, M3, and M4 are shown in Figure R1. The diagrams compare the mean predicted probabilities with the corresponding observed frequencies of positive outcomes across 0.1-width probability bins. The numbers associated with each point indicate the number of observations within each bin. Overall, all four models tend to overestimate the probability of the positive class, as the points lie below the identity line. For M1, predicted probabilities between approximately 0.5 and 0.7 correspond to observed frequencies of around 0.4, indicating overestimation in this range. M2 shows improved behavior compared with M1, although some overestimation remains, particularly at higher predicted probabilities, where predicted values of 0.7–0.8 correspond to observed frequencies of approximately 0.5. M3 and M4 exhibit a similar tendency to overestimate probabilities while producing predictions over a wider probability range, including values above 0.8, which are not assigned by M1 or M2. Overall, our selected model (M2) shows reasonably good behavior despite a tendency to overestimate probabilities at the highest predicted values. However, the highest-probability bin contains only 21 observations, so this estimate should be interpreted with caution.*

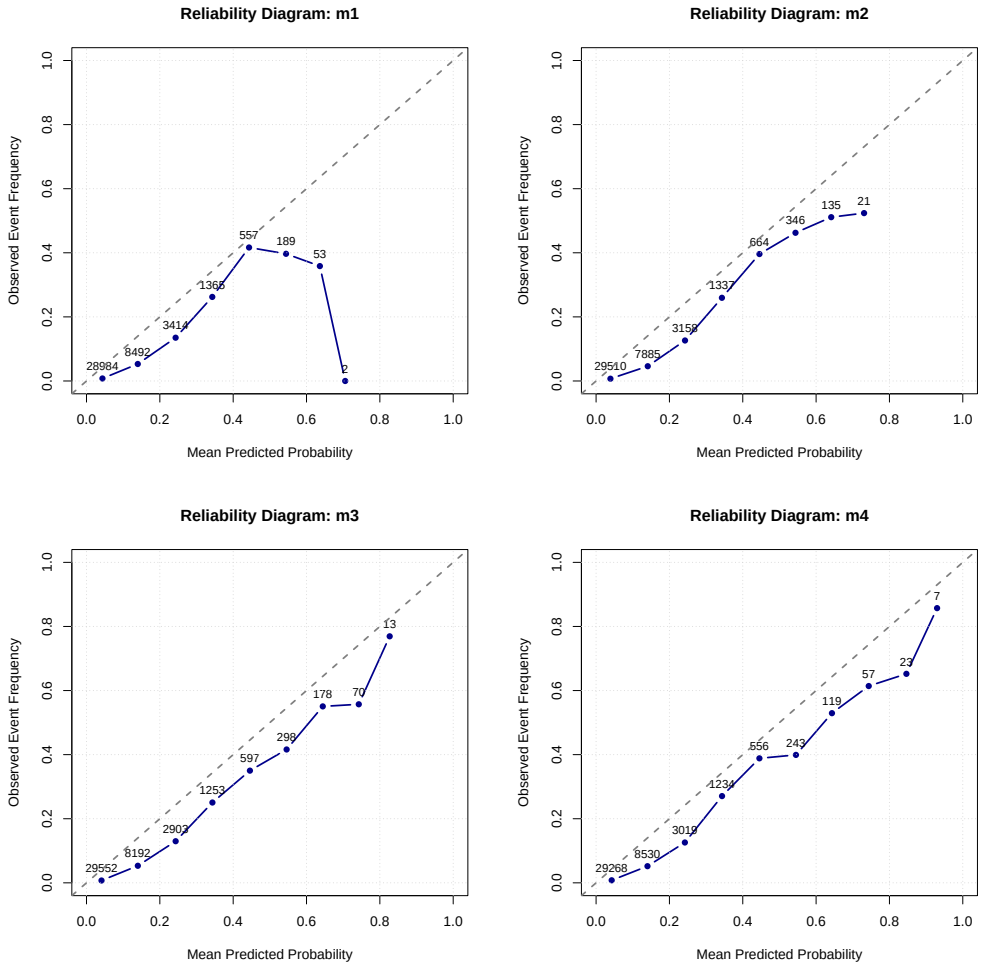

Figure R1: Reliability diagrams for models M1, M2, M3 and M4.

- *Calibration slope/intercept: The calibration intercepts obtained for all models are negative. On the other hand, the calibration slopes range between 1.607 and 1.645. Because these values are greater than 1.0, and the logits of the estimated probabilities are negative, so they indicate that the models overestimate the occurrence of record events during the validation period. For example, for M2, we found that the mean difference between the original predicted probabilities and the calibrated probabilities was 0.0699. Thus, the calibration reduces the predicted probabilities, indicating that the original model overestimated these probabilities.*

- *Sensitivity/Specificity at selected thresh-olds: By explicitly testing operational thresholds of 0.1 and 0.2, we show that at a 0.1 threshold, the models capture up to 89% of true events (Sensitivity) with a 71% Specificity. At a 0.2 threshold, the models capture up to 69% of true events (Sensitivity) with a 90% Specificity*

The results for the PR_{AUC} and the Brier Score will be included in Table 1 of the manuscript, as we consider them to be the most useful metrics for our problem. Table R1 will be included in the Supplementary Material, and these metrics will be described in Section 4.2 (starting at line 263). The paragraph will read as follows: "The models are compared according to the following criteria: (1) model parsimony (a low number of predictors); (2) Akaike Information Criterion (AIC), computed over the estimation period; and, on the validation period, (3) global AUC; (4) AUC performance for coastal stations (those located within 50 km of the se); (5) PR-AUC; and (6) the Brier Score. This multi-faceted evaluation (…)".

6. The proposed local-to-global selection procedure is interesting, but it relies on stepwise regression, z-score thresholds, and AIC-like penalties. Stepwise procedures can be unstable with correlated atmospheric predictors and spatially/temporally dependent observations. Please add a stability analysis, for example using block bootstrapping by year or leave-one-year/block-out resampling. Report how often each predictor is selected and provide uncertainty intervals for key model coefficients or partial effects.

   Response: We thank the reviewer for this constructive suggestion. We completely agree that spatial and temporal dependencies in atmospheric predictors can introduce instability in stepwise selection procedures. To address this concern and rigorously test the stability of our local-to-global selection framework, we performed a Leave-One-Year-Out (LOYO) resampling analysis across the years of our training period. This approach directly accounts for temporal dependencies by using annual blocks, by removing one full year of data at a time, and re-running the entire selection procedure on the remaining 49 years (50 total iterations). The selection frequencies for the predictors that passed the empirical filter are presented in Table R3. The results demonstrate remarkable stability in our selection procedure:

   - *15 predictors were selected in 100% of the resampling iterations.*
   - *20 out of 22 predictors maintained a selection frequency above or equal to 88%.*

   Only one predictor (g700.35N.10W) was selected in fewer than 50% of the LOYO iterations. As shown in Figure S.5 of the Supplementary Material, this predictor exhibits z-scores close to zero at many stations, indicating weak evidence for its inclusion.

Moreover, although this predictor passed the filter in Step 2 of the algorithm, it was discarded in Step 3 during the stepwise selection used to obtain model M2 (see Section S.2.3 of the Supplementary Material). Therefore, it was not included in the final model. Overall, these results indicate that the local-to-global selection strategy is robust, with only a single predictor showing a selection frequency below 50%.

Table R3: Selection frequencies of the local atmospheric predictors under Leave-One-Year-Out (LOYO) stability analysis.

Predictor	Freq.
g700..lag1	1.00
g500.45N.10W.lag1	1.00
g700.45N.10W.lag1	1.00
g300.35N.10W.lag1	1.00
g500.45N.5E.lag1	1.00
g700.45N.5E.lag1	1.00
g300.35N.5E.lag1	1.00
g700.	1.00
poly(g700.45N.10W, 2)1	1.00
poly(g700.45N.10W, 2)2	1.00
g300.35N.10W	1.00
g500.45N.5E	1.00
poly(g700.45N.5E, 2)1	1.00
poly(g700.45N.5E, 2)2	0.98
poly(g300.35N.5E, 2)1	0.98
poly(g300.35N.5E, 2)2	1.00
poly(g700.35N.5E, 2)2	1.00
g500.45N.10W	0.92
g500.35N.10W	0.92
g500.35N.5E	0.88
g500..lag1	0.70
g700.35N.10W	0.34

In the revised manuscript, Table R3 will be included in the Supplementary Material, and an explanation will be added to the paragraph in Section 5.1.2 (line 307). The paragraph will read as follows:

"Using a threshold for absolute z-scores greater than or equal to $z_{th} = 2$ in at least one-third of the sites (threshold $s_{th} = 1/3$ out of 36), $N_c = 23$ potential predictors for the

global model are selected. Table S.2 summarizes the number of predictors obtained under different combinations of thresholds z_{th} and s_{th}. ***To assess the stability of this selection filter, we performed a Leave-One-Year-Out (LOYO) resampling analysis across the years of the training period by removing one complete year of data at a time and repeating the entire selection procedure on the remaining 49 years (50 iterations in total). Table S.? presents the selection frequencies of the selected predictors. Only one predictor (g700.35N.10W) was selected in fewer than 50% of the LOYO iterations. As shown in Figure S.5, this predictor exhibits z-scores close to zero at many stations, indicating weak evidence for its inclusion. Although it passes the empirical filter, it is discarded during the stepwise selection. Overall, these results indicate that the local-to-global selection strategy is robust, with only a single predictor showing a selection frequency below 50%.*** *After applying the stepwise procedure to this set of 23 potential predictors, a model including 17 predictors is obtained, see diagram in Figure 6 and S.6."*

7. The global model concatenates binary indicators across stations and days, but the manuscript itself shows strong temporal persistence and spatial co-occurrence. This dependence may affect standard errors, z-scores, and variable-selection thresholds. Please clarify whether standard errors are corrected for clustering by day, year, or station. If not, either add cluster-robust/block-bootstrap uncertainty estimates or state clearly that coefficient significance is used only heuristically for screening.

   Response: The standard errors are not corrected. We agree with the reviewer that temporal persistence and spatial co-occurrence introduce dependencies that affect standard errors and affect z-scores. Because these temporal and spatial dependencies are not fully captured by the model, we agree that the standard errors and p-values cannot be interpreted as significance tests. Therefore, we will revise the manuscript to state explicitly that coefficient significance is used as a heuristic screening tool for variable selection, rather than for formal hypothesis testing. This will be included in line 246 and will read as follows: '· · · This guarantees cross-station representativeness. ***Note that coefficient significance is used as a heuristic screening tool for variable selection, rather than for formal hypothesis testing, as temporal persistence and spatial co-occurrence introduce dependencies that affect standard errors and z-scores.***"

8. The exploratory analysis shows strong persistence in record occurrence, and the manuscript evaluates whether simulated series reproduce run lengths. However, the fitted model does not appear to include an explicit previous-day record term or lagged Tx anomaly.

Please explain whether persistence is reproduced indirectly through geopotential persistence. A useful sensitivity test would compare the selected model with and without a previous-day record indicator or lagged Tx anomaly.

Response: We thank the reviewer for this insightful comment. Persistence in record occurrence is not included in the model through an explicit lagged term (e.g. previous-day record indicator or lagged Tx anomaly). Instead, it is reproduced indirectly through the temporal persistence of the large-scale atmospheric circulation, as represented by the lagged geopotentials. A substantial part of the linear predictor is associated with these previous-day terms, as the selected M2 model includes eight lagged geopotential predictors (Table S.2.3 of the Supplementary Material), two of which interact with longitude and two with latitude. These geopotentials evolve in time and reflect the synoptic conditions that govern successive days, therefore explaining part of the temporal dependence and introducing dependence in the simulated record sequences.

The lagged temperature terms provide proxy information about the previous day's atmospheric conditions, which are already represented in our model through the lagged geopotential predictors. In our framework, we aim to maintain a physically interpretable model driven by atmospheric covariates, rather than introducing purely statistical lagged response terms that may partially duplicate information already contained in the lagged geopotentials. To assess this, we evaluated our selected M2 model including the previous-day record indicator. This did not improve the discriminatory ability of the model, with the AUC in validation remaining at 0.89.

We acknowledge that alternative modeling strategies could include explicit autoregressive or Markov-type formulations based on previous-day response, but this is not our particular objective. Regarding the model's ability to reproduce run lengths, comparing our results with those of Chida et al. (2025), who also considered a setting in which the occurrences of 1s are sparse, we observe a similar difficulty in reproducing the run-length distributions (see Figure 6 in Chida et al. (2025)). In particular, the Hidden Markov Model substantially underestimates the occurrence of short runs. In our study, Table 2 also shows an underestimation of runs of length 2, although the observed frequency remains within the confidence interval.

9. Discuss spatial co-occurrence more cautiously: the model-simulated Jaccard indices show a positive relationship with the empirical Jaccard indices, but the simulated values are systematically lower in magnitude. Therefore, the model partially captures the spatial co-occurrence pattern but underestimates its strength. Please avoid wording suggesting that spatial dependence is fully reproduced.

   Response: You are correct that the wording in the conclusion section suggests that spa-

*tial dependence is fully reproduced. We will therefore revise the corresponding sentence in the conclusions section, which currently says: "Additional validation confirmed the model's ability to reproduce the observed persistence and spatial dependence of record events." We will replace it with: "**Additional validation confirmed the model's ability to reproduce the observed persistence and partially capture the spatial co-occurrence pattern, although its strength is underestimated.**"*

10. Provide more detail on data quality and station homogenization: please specify whether the ECA&D Tx series are homogenized, how missing values are handled in the record calculation, and whether urban heat island or station-change effects were screened.

    Response: We will clarify these aspects in the revised manuscript. Regarding the ECA&D series, we downloaded the blended dataset, which incorporates an automated update procedure based on daily SYNOP messages distributed in near real time through the Global Telecommunication System (GTS), together with additional gap filling using nearby stations.

    For the record calculation, missing values are treated as non-record days (assigned a value of 0). This has a negligible impact on our analysis because the amount of missing data is very small. Specifically, 25 stations have at most 10 missing values over the entire study period. The station with the largest number of missing values has 72 missing observations, corresponding to less than 1.5% of the observations.

    Urban heat island and station-change effects were not explicitly screened in our analysis beyond the quality-control procedures already implemented within the ECAD database. However, the stations used in our analysis are located at airports or in areas that are not affected by urbanization. The only exception is Madrid-Retiro, which is located in the city centre of Madrid but within a large urban park. We therefore consider that urban heat island effects are unlikely to have a substantial impact on our results.

11. The use of 12:00 geopotential at 1° resolution should be better justified. Since Tx often occurs in the afternoon, the authors should explain why 12:00 was selected instead of 00:00, 06:00, 18:00, daily mean fields, or geopotential-height anomalies. It would also be useful to clarify whether results change if geopotential is expressed as geopotential height in meters, which is more interpretable meteorologically.

    Response: The choice of the 12:00 UTC geopotentials was motivated by their representativeness of the large-scale synoptic conditions during the daytime period when maximum temperatures are typically occurring or have recently occurred over the Iberian Peninsula. We acknowledge that the use of a single analysis time represents a limitation of the present study, and that a more complete approach could consider other

synoptic times (e.g. 00:00, 06:00, 18:00 UTC), daily means, or anomaly fields, which may further increase robustness. This extension is part of ongoing and future work. In particular, we are interested in investigating the predictive ability of geopotential measures at 18:00 UTC.

Regarding the 1° *spatial resolution, this was selected as a balance between capturing synoptic-scale variability and maintaining robustness, given the strong spatial correlations of the geopotentials over the domain (Figure S.2 of the Supplementary Material). In addition, the closest meteorological stations are approximately 100 km apart, which approximately corresponds to a* 1° *grid spacing. Therefore, using a finer spatial resolution would not substantially modify the large-scale circulation patterns relevant to our application.*

Finally, geopotential and geopotential height are linearly related through a constant scaling factor, and therefore the use of one or the other does not affect the modeling results, only their physical units and interpretation. We used geopotential as provided directly by the reanalysis dataset to avoid unnecessary transformations and to remain consistent with the original data product.

12. Clarify the spatial prediction maps.

    The paper presents maps of predicted record probability during the August 2023 event. Please explain exactly how these maps are produced from a station-based model. Are predictions made only at station locations and interpolated, or are covariates and geodetic terms evaluated continuously over the grid? What climatic covariates are assigned to non-station grid cells?

    Response: We agree that the methodology used to construct these maps was not sufficiently detailed in the manuscript. Predictions are not computed only at station locations and subsequently interpolated. Instead, the fitted model is evaluated directly at each grid point of the (11×16) *spatial grid. For each grid point, the climatic covariates are evaluated using the values corresponding to that grid location and the four grid corners. Unlike the station-based setting, where covariates are linked to the nearest grid point to a station, here the covariates are taken directly at the prediction grid point itself. Similarly, the latitude and longitude terms included in the model are evaluated using the coordinates of each grid point.*

    This procedure allows us to obtain estimates of record probability over the study domain for each day using the selected model. The resulting probabilities are then smoothed for visualization using kriging interpolation. Since the model was estimated exclusively

from observations at mainland Spanish stations, the interpolation is restricted to the territory of mainland Spain. We selected the August 2023 event because it corresponded to an exceptionally hot period with widespread temperature records, providing a relevant case study to assess the model performance under extreme conditions.

As noted by the reviewer, the model could obtain probability estimates without the need for interpolation by evaluating the covariates and geodetic terms continuously over the grid.

To clarify this, we will add an explanation to the paragraph beginning on line 363, which will read as follows: "(…) each depicting the predicted probability of a record occurring on different days within a heatwave. ***To obtain these maps, the fitted model is evaluated directly at each grid point of the spatial grid. Unlike the station-based setting, where covariates are linked to the nearest grid point to a station, here the covariates are taken directly at the prediction grid point itself. Similarly, the latitude and longitude terms included in the model are evaluated using the coordinates of each grid point. This procedure allows us to obtain estimates of the probability of record events over the study domain, which are then smoothed for visualization using kriging interpolation. Since the model was estimated exclusively from observations at mainland Spanish stations, the interpolation is restricted to the territory of mainland Spain. Alternatively, the model could be evaluated on a finer prediction grid, allowing record probabilities to be obtained directly at a higher spatial resolution, without the need for the kriging interpolation.*** *These visualizations highlight the model's (…)"*

Finally, we would like to thank the reviewer for the insightful, constructive, and helpful comments which have improved the quality of the manuscript. We also thank the reviewer for the careful technical corrections and for identifying the typos and errors throughout the manuscript. All of these will be corrected in the revised version.

References

Chida, T. E., Ramesh, N., Onof, C. & Yip, I. (2025), 'Spatial and Temporal Dependence Framework in Multi-Site Precipitation Modelling', *Stochastic Environmental Research and Risk Assessment* **39**(9), 3781–3811.