Response to Anonymous Referee #1 on Manuscript ID EGUSPHERE-2026-832

Anonymous Referee #1:

This may be an interesting study, but I found the text a bit difficult to digest due to vague formulations and long and windy descriptions of aspects that I didn't see a clear connection to the core of this topic which this paper is about. I'm not sure if I even quite understood this core topic. Did the authors use AI/ML in assisting them to read this? Also, the paper lacks proper acknowledgement and citation of earlier work related to this study. Finally, I found the paper a bit long.

Response: We sincerely thank the reviewer for the time and effort spent reviewing our manuscript and for considering it a potentially interesting study. We have carefully considered all of the comments and suggestions. We note that the reviewer mentions having stopped reading the manuscript on page 15. As a result, several of the questions and concerns raised in the review are addressed in the later sections of the manuscript, particularly in the Results, Discussion, and Conclusions. We believe that a complete assessment of the study requires reading the full manuscript. Finally, we would like to clarify that no AI tools were used to write or generate this manuscript, all of the work and writing are entirely original. We therefore begin by clarifying the objective of the study and our background in this field before addressing the reviewer's specific comments.

The objective of this study is to develop a methodological framework for modeling the occurrence of calendar-day Tx records and their relationship with atmospheric geopotentials using data from multiple stations within a region. The proposed methodology is then applied to a Spanish database to characterize the occurrence of record-breaking events and compare the observed behavior with that expected under stationarity. Our research group has extensive experience in the development of statistical methods for climate applications, including statistical modeling using atmospheric covariates, the analysis of climate extremes, statistical downscaling of temperature and precipitation, and the statistical study of record-breaking events. In this work, however, our objective is not to develop or extend statistical theory, but to provide a general statistical framework that can be applied across a wide range of climate studies.

1. The abstract is a bit hard to follow. Please consider rephrasing so that it's easier to understand for a little more general readership. E.g. I'm not sure what the meaning is of "Among the tested models, a global model that enhances the base model with geodetic interactions was selected for optimal balance between predictive performance and complexity"

*Response: We appreciate the reviewer's comment. In the abstract, our intention was to summarize the statistical modeling strategy developed in this work, particularly the selection of a global model that incorporates interactions between atmospheric covariates and geodetic variables describing the station location. Nevertheless, we agree that the original wording could be made clearer, and we will therefore revise the sentence so it reads as follows: "**Among the tested models, a global model extending the base model by including interactions with geodetic variables describing the station locations was selected for its optimal balance between predictive performance and complexity.**"*

2. The introduction does not acknowledge the early works that applied record-breaking statistics in climate research. See the overview provided by e.g. DOI:10.1126/sciadv.ado3712. Also, there is no mention of independent and identical distributed (iid) data in the introduction, which the authors seem to label 'stationary climate'. And the claim "... their occurrence follows well-established, distribution-free probabilistic properties" is not quite correct, as the records follow certain patterns as described in previous (uncited) studies. There is also a possibility that autocorrelation may introduce a higher recurrence of record-breaking events. While autocorrelation hardly seems to be an issue for calendar-date record-statistics (but adjacent days may be associated with annual series which are not entirely independent - so it may be more clean to subsample the days - eg every 5 day), it is nevertheless important to discuss the meaning of iid for completeness in he introduction and to convince the reader that they actually are familiar with record-breaking statistics.

   As I read the introduction now, it does not provide a convincing argument for the authors having a good knowledge of this topic beyond citing AEMET and NOAA, and that also makes me more doubtful about the rest of their analysis. The neglect to acknowledge previous efforts relevant for the topic can be a slippery slope and cause part of the understanding to fade away when others in turn cite papers with incomplete background information. But to be fair, the concept of iid is mentioned later on in the paper under section 3.1.1, but not properly discussed and it's nevertheless good to introduce it in the introduction.

   Response: The introduction is designed with the intention of motivating two main aspects of this work: (1) the use of record-breaking events as an indicator of climate change, using the stationary (iid) case as a reference, and (2) the need for new statistical methods capable of characterizing the relationship between atmospheric variables and the occurrence of records. Therefore, the introduction focuses on motivating the problem addressed in this paper rather than on the theoretical background of record-breaking

statistics.

We acknowledge the valuable contributions of the pioneering authors who established the theoretical foundations of record-breaking statistics. To better reflect this background, we will include the corresponding references in line 35. The sentence will read: "…well-established, distribution-free probabilistic properties (Benestad 2003, 2004, Arnold et al. 1998, Diersen & Trenkler 1996, Foster & Stuart 1954)."

We chose this structure because we wanted the introduction to focus on the relationship between record-breaking events and the atmospheric situation. We believe this organization allows the introduction to better motivate the objectives of the study.

3. I expect there will be a connection between geopotential covariates and heatwaves, but it's not so clear for record-breaking events. Heatwaves and record-breaking events have different definitions and different (statistical) behaviors. Record-breaking events have a simple and clear definition, whereas the definition of heatwaves varies from country to country, typically being a number of consecutive days with temperatures above some threshold. A record-breaking event depends on the number of previous measurements and their measured values, and involves a degree of randomness connected to the time and place of the data sampling. This randomness is expected to make a connection more tricky, while I also am convinced that records tend to fall during blocking highs.

   Response: We share the reviewer's curiosity regarding the possible relationship between record-breaking events and the atmospheric situation, which the reviewer considers more evident when extremes are defined as heatwaves. Indeed, the fact that this relationship is not obvious is exactly what motivates this study. The objective of this work is to determine whether atmospheric geopotentials contain useful information for characterizing the occurrence of record-breaking events, despite their probabilistic nature. Precisely to address this question, we developed the methodological framework presented in this paper.

   Our final model, based on atmospheric covariates, achieves an AUC of 0.88 on the validation period. We therefore believe that the results presented in this work provide strong evidence of a meaningful relationship between atmospheric geopotentials and the occurrence of record-breaking events.

   This issue is also discussed in our response to Anonymous Referee #2 (Comment 3), where we compare our proposed model with a stationary baseline model, a time-trend-only model, and a detrended model.

4. It's a bit annoying that some vague statements about bi-variate analysis or specific downscaling are left in the air and the reader is left to guess or confer with Castillo-

Mateo et al. (2025a,b). Please provide a clear and exact explanation so that I don't have to take a break and start reading another paper to understand what the deal is. Furthermore, the cited period 2012–2021 is a fairly short time period (10 years?), and one question is whether 10 data points are sufficient for a proper iid-test - but this is after all a cited study and perhaps not so relevant for this? I'd expect this short time period to be influenced by decadal variability, and would reserve the concept trend to longer time spans. Why this short time period in the first place? Anyway, this part brings me some doubts about whether we have a common understanding of record-breaking statistics, and more the reason for providing a proper account and background to the matter.

Response: The questions raised by the reviewer refer to the methodology of a cited paper rather than to the present work. Our intention was to cite Castillo-Mateo et al. (2025a,b) as background, not to reproduce or discuss all the methodological details of the study. Nevertheless, we clarify this point below.

Regarding the validation period, note that our validation period covers 2011 to 2023. This period was selected because it corresponds to the most recent years, where there is clear evidence of deviations from the iid assumption, thus providing a particularly challenging validation of the proposed model.

The cited study also uses a relatively short validation period of 10 years, while ours contains 13 years. However, the analysis is not based on only 10 observations in their case or 13 in ours. Instead, their validation dataset comprises 122 observations per year, whereas ours comprises 92. In both studies, the objective is to estimate the probability of a binary outcome over the complete validation dataset.

5. I'm not sure that geopotential covariates are well-suited for developing statistical downscaling methods to project the future frequency of record-breaking temperatures under global warming scenarios because it's not clear whether such predictors embed the climate change signal that well. This is the ultimate purpose of this study, is it not? Another alternative could be to downscale the shape of pdfs for the temperatures to infer how their statistics change. A stretching of the tails of the pdfs, after all, provides information about probabilities of record-breaking temperatures. Furthermore, the iid-test was originally inspired by the question whether these distributions were non-stationary. One suboptimal strategy is nevertheless to downscale each data point as in the traditional ways because such an approach would miss the extremes, which I think this paper tries to allude to.

   Response: This is a very interesting question. Anonymous Referee #2 raises a similar question in Comment 3, where we address this issue in detail. In particular, we compare

the proposed model with a stationary baseline model, a temporal trend-only model, and a version of the selected model fitted using detrended geopotentials. The comparison with the detrended model allows us to assess whether the predictive ability of the proposed model can be attributed only to the long-term warming trend or whether large-scale atmospheric circulation also explains part of the variability in record occurrence, and our results show that it does. For more details, we refer the reviewer to that response.

6. It may be more correct to say that the probability for a new record an iid variable is the reciprocal of the number of observations (n), as the description provided doesn't specify when time t starts. It could be time if t is the number of time steps (t is a discrete number and not a rational number in this case).

   Response: In our notation, t denotes the year index, taking values from 1 to 64 (discrete numbers), rather than the number of observations. Calendar-day index is denoted as ℓ, and records are defined separately for each day of the summer season by comparing the temperature on a given calendar day with the temperatures observed on that same calendar day in all previous years. Consequently, we have 92 calendar-day series, each consisting of 64 yearly observations. Therefore, the probability of a record in year t is $1/t$, where t refers to the year index within each calendar-day series.

7. Please explain how equation 3 was derived or where it comes from.

   Response: Equation 3 is not derived in this work but corresponds to the standard definition of the log odds ratio. In our work, it is used to quantify the effect of memory on the odds of record occurrence by comparing the odds of observing a record when the previous day was also a record with the odds when the previous day was not a record.

8. In Fig3b, should I not expect the iid case (red curve) to lie over the empirical counts (blue curve) since there are more observed records than for the iid case? The y-axis is after all probability of non-records. I'm confused. Please explain more carefully.

   Response: This figure is devoted to illustrating the strong spatial dependence of record occurrence in our study region, i.e., their co-occurrence. Although the observed series contains more records than expected under the iid assumption, the probability of observing no records at any station is still higher than under independence. This is because record-breaking events tend to occur simultaneously across stations due to the common atmospheric situation, rather than independently. Consequently, non-records also tend to occur simultaneously, increasing the probability of observing all zeros compared with the iid case.

   In particular, Figure 3b shows in blue the empirical relative frequency of observing no records at any station in the study region. As a reference, the red curve corresponds to

the probability of zero successes in a binomial distribution with n_S stations and success probability $1/t$, representing the iid case. The observed evolution shows that there is no clear convergence to 1, highlighting the deviation from the stationary (iid) case.

9. Need a proper reference for the Jaccard index. It's also appropriate to acknowledge others who have contributed with such indices and knowledge.

   Response: We will modified lines 166–167 to include the reference Real & Vargas (1996). The text will read: "Jaccard index. Another useful way to quantify co-occurrence between two series of records is by using the Jaccard index, a similarity coefficient for binary variables (Real & Vargas 1996)."

10. G300 is not properly introduced and it's not clear. Also "Among the geopotential variables, G300 shows the most similar evolution to Tx, and the trend appears to weaken with increasing height" is confusing. Isn't G300 the geopotential highest up?

    Response: The reviewer is indeed right that G300 corresponds to the highest geopotential level. This was a mistake on our part. We will rephrase the sentence in lines 189–190 to clarify this point. It will read: "Among the geopotential variables, G300 shows the most similar evolution to T_x, and the trend appears to weaken with ***decreasing*** *height."*

11. For record-breaking statistics, I'd expect to see similar shapes in the cumulated sums of events, as the four panels seem to demonstrate. I cannot see the curve for the iid case in Fig.4, and the figure caption is not very helpful.

    Response: The iid case is represented by the gray line. Since it is thinner than the other curves and, for example, overlaps with the observed points in the Huelva panel, it may be difficult to distinguish. However, it is clearly visible in several parts of the other panels, such as the right-hand side of the Barcelona-Fabra and Murcia panels and the central part of the Coruña panel. We hope this clarification helps the reviewer identify it.

12. It's not helpful to cite figures in the supporting documents, which means that I need to jump between different documents.

    Response: We are sorry that the reviewer found this inconvenient. However, due to the page limit, some figures had to be moved to the Supplementary Material. We believe this is an appropriate way to provide additional supporting information while keeping the manuscript within the journal's length requirements.

13. It may be useful to consider the model's minimum skilful scale for using the output as a predictor for downscaling. The models' minimum skilful scale is not well defined,

but there have been suggestions that it's about 4x4 grid boxes (DOI:0.2151/jmsj.2015-042). One may justify the choice based on that, which is relevant for when this method is applied to the output for global climate models.

Response: Thank you for your valuable suggestion. Anonymous Referee #2 raises a similar comment (Comment 11), where we justify the use of the 1° *spatial resolution. In particular, we explain that this resolution was selected as a balance between capturing synoptic-scale variability and maintaining robustness, given the strong spatial correlation of the geopotentials over the study domain. Moreover, the closest meteorological stations are approximately 100 km apart, which roughly corresponds to a* 1° *grid spacing. Therefore, we believe that this spatial resolution is appropriate for our application and refer the reviewer to that discussion.*

14. The question of how many tests were carried out should also be reflected in the estimation of the so-called 'field significance' (DOI:10.1175/JAM2404.1). When there are many combinations, then I'd expect that one or two will score high based on just chance.

    Response: Although our variable selection procedure uses the magnitude of the estimated z-scores as a filtering criterion, we do not perform a multiple hypothesis testing procedure. The z-scores are used to identify relevant predictors across stations within our local-to-global selection framework, rather than to perform formal significance tests. To assess the robustness of the selected predictors, we additionally performed the Leave-One-Year-Out (LOYO) resampling analysis described in our response to Anonymous Referee #2 (Comment 6), which showed that the selected variables are highly stable across resampling iterations.

15. Fig. 5 is complex/difficult and not easy to understand.

    Response: Figure 5 is intended as a summary of the methodology proposed in the paper, with each step also being explained in detail in the text. Its purpose is to provide a visual summary of the methodological workflow, illustrating each step of the statistical modeling approach.

16. L291. How was random chance quantified?

    Response: By "random chance"' we refer to the performance of a classifier with no discriminatory ability, corresponding to an AUC of 0.5. Since all fitted models achieve substantially higher AUC values, they all outperform random classification.

17. Exactly what does "a positive effect on the probability of record" mean?

Response: A positive effect on the probability of record means that the corresponding regression coefficient is positive. Therefore, as the value of that geopotential covariate increases, the estimated probability of observing a record-breaking event also increases, while a negative coefficient indicates the opposite.

Finally, we would like to thank the reviewer for all the insightful comments, which have helped improve our manuscript.

References

Arnold, B., Balakrishnan, N. & Nagaraja, H. (1998), *Wiley Series in Probability and Statistics*, John Wiley Sons, Ltd, pp. 313–319.

Benestad, R. (2003), 'How often can we expect a record event?', *Climate research* **25**(1), 3–13.

Benestad, R. (2004), 'Record-values, nonstationarity tests and extreme value distributions', *Global and Planetary Change* **44**(1), 11–26. Extreme climatic events.

Diersen, J. & Trenkler, G. (1996), 'Records Tests for Trend in Location', *Statistics* **28**(1), 1–12.

Foster, F. G. & Stuart, A. (1954), 'Distribution-Free Tests in Time-Series Based on the Breaking of Records', *Journal of the Royal Statistical Society. Series B (Methodological)* **16**(1), 1–22.

Real, R. & Vargas, J. M. (1996), 'The probabilistic basis of Jaccard's index of similarity', *Systematic biology* **45**(3), 380–385.