Response to reviewers

We are pleased to see that the reviewers value the content of our study. We appreciate their feedback and suggestions. Below, we provide a detailed, point-by-point response to the comments from the reviewers.

Our responses to reviewer comments are organized by category. Each response is labeled with a code in the specified range. The response categories are given below.

Reviewer Comments	Author Responses
CC1	A1-A2
RC1	B1-B26
RC2	C1-C30
CC2	D1-D5

RC2:

The paper shows an interesting study that teases out how different potential drivers for extreme heat emerge at time scales. The shorter (1-day) time scales are effectively a proxy of the "initial condition" forecast problem, from the atmospheric perspective, in weather prediction, and the strong role of circulation features bears that out. At longer time scales, the "boundary conditions" (i.e., land surface) emerges as an important factor. It brings to mind the point being made in the "infamous" figure used widely in the subseasonal-to-seasonal community (https://www.weather.gov/sti/stimodeling_s2sreport).

The main weakness of the manuscript is a lack of sufficient detail in the description of the methods - I believe this can be easily addressed. The main weakness of the study, as it weakens the conclusions, is the lack of significance testing of the trend analysis. I realize it is not applicable to all the methods shown, but certainly the part of the research comparing changes from the first to second decade of this century could be tested (see specific comments). Otherwise, I think the study has strong merit, and the manuscript can be published after some revisions described below.

C1: We thank Paul Dirmeyer for highlighting the merit of our study, and for the constructive comments below.

General comments:

1. An idea that emerges from this work is validation of the long-held notion that it is circulation features such as stationary ridges that initiate heatwaves (this is clearly stated in a couple of places), but that the land-atmosphere feedbacks (via surface drying and warming – much work by D. Miralles and colleagues on this) can both amplify and prolong heatwaves. The second aspect, prolonging heatwaves, is particularly well demonstrated by this study, and should be emphasized more in the abstract and conclusions, in my opinion. It comes from the novelty of the way a range of timescales has been investigated.

C2: We will include the paragraph below in the conclusion section to highlight the novelty of timescales:

"We reveal that land-atmosphere feedbacks substantially amplify and prolong these events as also shown by Miralles et al (2014). By examining both 1-day and 7-day timescales, we capture different phases of heatwaves—1-day events reflecting the peak of extreme heat, while 7-day events represent prolonged conditions. This approach allows us to infer that atmospheric drivers are likely more relevant in the intensity, whereas land surface drivers, such as surface drying and reduced evaporative cooling, become increasingly important as hot extremes persist."

- Miralles, D. G., Teuling, A. J., van Heerwaarden, C. C., & Vilà-Guerau de Arellano, J. (2014). Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. *Nature Geoscience*, *7*(5), 345–349. https://doi.org/10.1038/ngeo2141

We have also updated the sentence in the abstract about the increasing relevance of land surface drivers from daily to weekly time scale as follows:
"The relevance of land surface drivers increases from daily to weekly time scales, supporting the notion that heatwaves are prolonged by land-atmosphere interactions after they are introduced by the atmospheric circulation."

2. Another conclusion that I reached from reading this paper, based on the clear role of EVI (and EF, which is related to canopy conductance that itself links to vegetation carbon uptake and plant processes that regulate that), but that is not made by the authors, is that the results advocate for the inclusion of vegetation phenology in forecast models of weather and subseasonal climate. It is a bit "connecting the dots", but these relationships are arising from processes that are not a part of any operational forecast model (i.e., not parameterized in their land surface schemes), and are even absent from many CMIP models. In the final paragraph of the conclusions, you should point to operational prediction models specifically.

C3: Thank you for your feedback. We will add the modified paragraph to the conclusion.

"….This finding suggests that inclusion of vegetation phenology in operational weather and subseasonal climate forecast models could be crucial, as variables like EVI and EF are linked to vegetation processes, such as canopy conductance and stomatal resistance, which play a significant role in driving hot extremes. Many current forecasting models do not sufficiently exploit the available vegetation data such that they e.g. use only mean seasonal cycles instead of near-real time dynamics. Including these processes would improve the representation of land-atmosphere interactions, which is vital for enhancing the accuracy of hot extreme predictions."

3. Regarding land surface drivers for hot extremes, the literature review is quite short and Euro-centric for a paper with "global" in the title. There are other highly relevant citations in the recent literature that should be noted; a few I am quite familiar with:

https://doi.org/10.1029/2020AV000283, https://doi.org/10.1175/JCLI-D-20-0440.1, https://doi.org/10.1175/JCLI-D-22-0447.1, https://doi.org/10.1029/2023WR036490.

C4: We thank the reviewer for pointing us to these references. Here we give a section of the introduction where we emphasize the land surface drivers for hot extremes with more details.

"….On the other hand, land surface feedback mechanisms, including evaporative cooling deficits and vegetation water stress due to low soil moisture can exacerbate the hot extremes and lead to multi-hazard events (Wulff & Domeisen, 2019, Teuling et al., 2010, Miralles et al., 2014, Hauser et al., 2016). Similarly, a study by Benson & Dirmeyer (2021) identified a critical "soil moisture breakpoint," below which the probability of heatwaves increases due to a shift in surface energy fluxes from latent to sensible heat. This sensitivity becomes even more pronounced as soil moisture approaches the "permanent wilting point," where vegetation can no longer draw water from the soil, leading to a substantial increase in local surface temperatures. As a result, the sensitivity to soil moisture deficits significantly contributes to the severity of heat events (Dirmeyer et al., 2021). This effect underscores the spatial variability of soil moisture–temperature feedback mechanisms across different climatic zones. Specifically, transitional regions where latent heat flux strongly depends on soil moisture, exhibit more pronounced land-atmosphere coupling (Wehrli et al., 2019; Koster, 2004)...."

Specific comments:

1. L65: It took a while for me to realize that by "height differences" you mean horizontal gradients, relevant to the geostrophic wind relationship. In atmospheric thermodynamics, the term "height differences" is typically applied with respect to the hypsometric relationship, i.e., the vertical distance or "thickness" between two pressure levels, which relates to the mean virtual temperature of the layer between. To avoid confusion, you should replace "height differences" with "horizontal height gradients", or just be explicit that this is a proxy for the geostrophic wind (you could label this as "Advection" here and in Figures 2, 5, A1, A6, A7).

C5: We agree that 'height differences' may lead to confusion. We have replaced 'height differences' with 'horizontal height gradients' in L65.

"In addition, we compute the horizontal geopotential height differences at 500 hPa pressure level for each grid cell with respect to the values in adjacent grid cells in the northern, eastern, southern and western directions."

2. §2.1: It is stated that daily data (shortest time scale) are used. Are these based at each point on the local time, or all on 0000UTC as the day boundary? If the latter, then for about half of the world, what you call "one day" actually spans two days with respect to important diurnal phase of drivers like net radiation and evaporative fraction. Please clarify and/or justify the choice.

C6: Thank you for the valuable feedback. We confirm that the ERA5 daily means used in our analysis are computed based on UTC+00:00. We acknowledge that this approach may lead to phase mismatches in diurnal cycles for some variables, particularly in regions where local time differs significantly from UTC (Zou and Qin, 2010). However, we believe this does not substantially affect our results, as the drivers in our analysis do not change significantly from one day to the next, and the diurnal cycle is inherently accounted for when using daily means.

We have clarified this in L51:
"The daily means used in our analysis are computed based on UTC+00:00. While this choice may lead to phase mismatches in diurnal cycles for some variables, particularly in regions where local time differs significantly from UTC, it provides consistency across datasets, which is essential for our analysis."

- Zou, X., & Qin, Z.-K. (2010). Time zone dependence of diurnal cycle errors in surface temperature analyses. Monthly Weather Review, 138(6), 2469–2475. https://doi.org/10.1175/2010mwr3248.1

3. L70-71: This is the first time either "X-BASE" or "ERA5" are mentioned (before Table 1 is cited). They should be defined or described here, or else moved after Table 1.

C7: Here's the revised sentence in L70-71:
"It is important to note that we compute EF using variables from two different datasets: X-BASE and ERA5. This approach is justified, as X-BASE is formulated using ERA5 data."

4. Table 1, EF: How is EF calculated from X-BASE? To my knowledge, the publicly released data does not contain this variable, nor the necessary data to calculate it (i.e., there is no sensible heat flux field). That renders this part of the study unreproducible by others.

C8: Thank you for your question. The relevant calculation is explained in line 69 (section 2.1):

"EF is computed by normalizing evapotranspiration (ET), which we retrieve from X-BASE, by surface net radiation, which we retrieve from ERA5."

5. L92: How are the warm seasons defined? There are a number of approaches, as there are strong latitudinal (and more complex) determinants. Are the same number of months used everywhere, or are only very cold months avoided (a temperature threshold)?

C9: The warm season for each grid cell is defined by identifying the hottest day based on absolute temperature and then applying a 60-day window centered around that day. This approach is repeated for all years within the study period, ensuring consistency in the number of days considered as the warm season for each grid cell. Consequently, each grid cell has the same number of days defined as the warm season across all years, based on local temperature patterns.

6. Figure 1: This is an important figure, but it not clear and the descriptions in the text do not fully clarify the workflow, especially for part (c). In the caption, it should explicitly say “see text for details”.

C10: Thank you for your comment on Figure 1. We appreciate your feedback, and since similar points were raised by the other reviewer (see response B10), we have made adjustments to improve the clarity of the figure and the workflow description, especially for part (c). The revised figure and accompanying text now provide a clearer explanation of the workflow. We have also included 'see text for details' in the caption as suggested.

7. L100: Please give more description of the definition of “similar” (i.e., please do not rely solely on a reference to Yiou et al. 2007). Is it based on RMSE? Is some normalization applied? Perhaps it is best to include equations.

C11: We identify analogue periods by selecting the five periods with driver values most comparable to those observed on the hot days based on one dimensional euclidean distance. We have added additional explanation to the manuscript to describe our selection process in detail. This clarification has been added to section 2.3.

“....This means that for each driver and at each considered atmospheric level (i.e., geopotential height and wind) and temporal scale (i.e., EVI, EF, and surface net radiation) we select the five periods with the raw driver values most similar to those observed during the identified hot extremes based on one dimensional euclidean distance. This approach shares a conceptual basis with the analogue methods in the literature, such as those used by Jézéquel et al. (2018) and Yiou et al. (2007). These studies show that selecting more than five geopotential height analogues has little significant effect on the results. Also to maintain consistency across all grid cells, we use the same number of analogues in our analysis.”

- Jézéquel, A., Yiou, P. & Radanovics, S., (2018): Role of circulation in European heatwaves using flow analogues. Clim Dyn 50, 1145–1159 . https://doi.org/10.1007/s00382-017-3667-0
- Yiou, P., Vautard, R., Naveau, P., & Cassou, C., (2007): Inconsistency between atmospheric dynamics and temperatures during the exceptional 2006/2007 fall/winter and recent warming in Europe. Geophys Res Lett, 34(21), https://doi.org/10.1029/2007gl031981

8. L103: This is not entirely clear – do you mean that the center of the window is on the calendar date (month and day) applied across all years?

C12: The ±60-day terminology we used in the text actually refers to a 120-day window around the specific calendar date (month and day) of the hot extreme event, applied across all years to select the analogue periods. This approach ensures that each analogue period is centered on the same seasonal timing as the event while maintaining at least 15 days of separation to ensure independence. We acknowledge that the phrasing is not clear here. We have revised the sentence in section 2.3.

"....For this purpose, a 120-day window surrounding the specific calendar date (i.e., month and day) of the relevant hot extreme event is considered across all years to select the analogue periods. These selected periods are also at least 15 days apart from each other to ensure independence."

9. §2.4: As noted above, this description is very fuzzy. I do not follow the process. Again, perhaps equations or pseudocode is needed. I would not be able to reproduce this methodology based on the description.

C13: We've added an equation in section 2.4 to better explain the methodology:

"For each variable *D* (e.g., geopotential height, EVI, EF), we identify 15 analogue periods based on similarity to the three hottest observed extremes (5 analogues each) in each grid cell. The degree of relevance for each variable "D" in each grid cell "g" is computed as follows:

$$Degree\ of\ relevance(g, D) = \frac{1}{15}\sum_{i=1}^{15}\left(\frac{T'_{analog}(g,i)}{\bar{T}'_{event}(g)}\right)$$

T'_analog(g, i) denotes the temperature anomaly during the i-th analogue period in grid cell g, based on the conditions of driver D.
T'_event(g): This is the mean temperature anomaly calculated from the three observed hottest extreme events in the grid cell g. It serves as a basis of comparison to determine how much of the observed extreme temperature anomaly can be explained by the analogue temperature conditions of variable D."

10. L114-115: This is, of course, a linear assumption, that the drivers can be considered separately. This point is acknowledged later as a possible drawback, but it would be good to state that here – this is where some readers will begin to have this question in their minds.

C14: We've added the following lines:
"It is important to note that this approach assumes a linear and separate contribution of each driver, which may be a limitation when interactions between drivers are relevant."

11. L120: Should "both" be replaced with "each of the"?

C15: We have changed the wording as you suggested:
" For this purpose, we divide the study period into two periods, 2001-2010 and 2011-2020, and employ the same methodology as described in Sections 2.1 to 2.4 to calculate the relevance of all driver variables for each of the time periods."

12. Figures 2, 4, A1, A5, A7: It is difficult to tell the grey from some of the pale blue shades – they have very similar luminance. Additionally, the monochrome palettes in Figures 4 and A5 make them somewhat hard to read. It appears that you are trying to be

considerate of colorblind readers – using a cubehelix palette in these two figures would improve clarity for all.

C16: We will change the background color to a darker gray to improve contrast with the pale blue shades. We will also use a cubehelix palette with varying luminance in Figures 4 and A5 to enhance readability and ensure accessibility for colorblind readers.

13. L142-143: Please move this final sentence of the paragraph up to become the 2nd sentence (right after Fig. 2 is mentioned).

C17: The sentence is now in line 128:
"The global distribution of the dominant variables for both 1-day and 7-day time scale extreme temperatures are illustrated within Fig. 2. A more detailed depiction of drivers' relevances across height levels and time scales is presented in Fig. A1."

14. L154-163: This is methodology: it should be explained in §2, not with the results. Additionally, how the Random Forest method is applied must be explained in sufficient detail such that a reader could hope to reproduce it.

C18: The relevant section will be moved to section 2.1 and updated as follows:
"In order to analyze the spatial distribution of the dominant driver variables identified for 1-day and 7-day hot extremes with respect to different land surface characteristics and climatic regimes, we employ a random forest approach where geopotential height and EF serves as target variables while a range of hydro-climatological, vegetation and landscape variables is used as predictors. The data were processed by separating the target variables from the predictors and splitting them into training and testing sets, with 25% of the data reserved for testing and a random state of 42 to control the shuffling applied to the data before applying the split. We used 100 trees, and a maximum depth of 10 to configure the RandomForestRegressor, as these hyperparameters have proved to work well in other studies (Oshiro et al.; 2012; Probst & Boulesteix 2017). Bootstrapping was enabled, and the feature importance was evaluated using SHAP (Shapley Additive Explanations) values to provide insight into the contribution of each predictor. The mean absolute SHAP values were calculated, and we found that long-term mean temperature and radiation are the most relevant predictor variables for both 1-day and 7-day hot extremes. Additionally, aridity (calculated as the ratio of long-term mean net radiation and unit-adjusted long-term mean precipitation) and topography play a role while the other considered variables are less important. While temperature has the highest relevance and is therefore selected as the primary variable, radiation, which ranks second in relevance, is closely related to temperature as an atmospheric variable. To ensure the inclusion of a land surface-related factor, we choose aridity, which captures the interaction between radiation and precipitation, thus providing a metric for assessing land surface influences on hot extremes. (Fig. A2)."

- Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest? In Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13–20, 2012, Proceedings, Vol. 7376, 154. Springer.

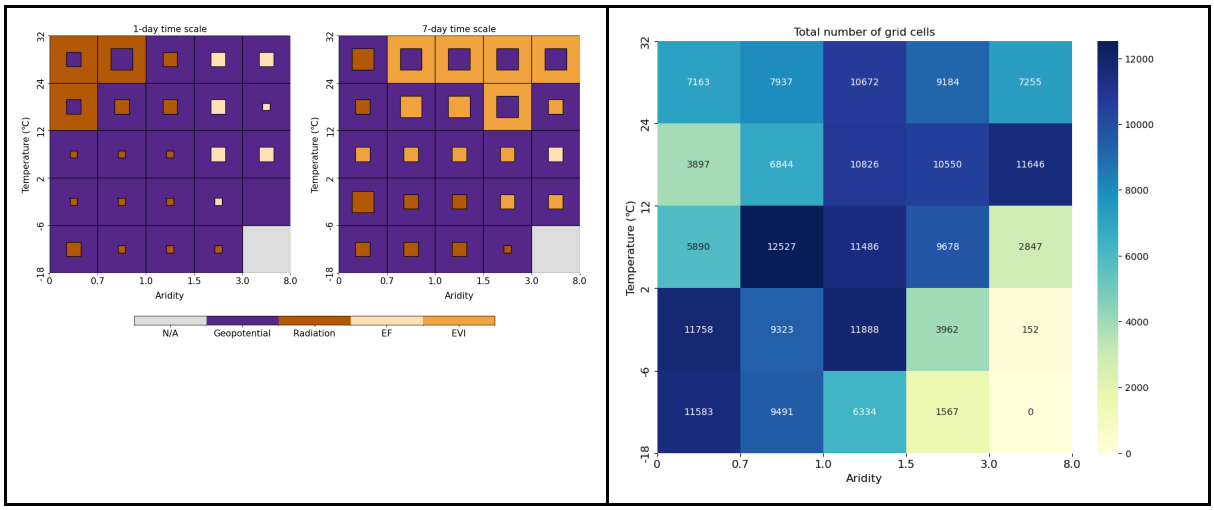
- Probst, P., & Boulesteix, A.-L. (2017). To tune or not to tune the number of trees in a random forest? Journal of Machine Learning Research, 18, 1–18.

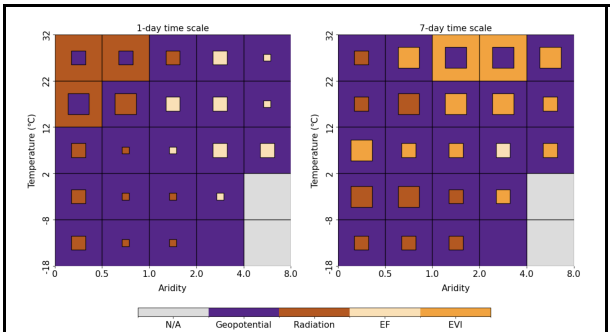
15. L168: Replace “mostly just” with “barely”.

C19: “Water availability is barely sufficient for vegetation in these regions, which means that (i) it can provide significant evaporative cooling; however, (ii) during warm and dry conditions, the limited water availability becomes insufficient, leading to reduced evaporative cooling and consequently enhanced temperatures.

16. Figures 3, A3, A4, A8: Aside from aridity=1.0, which has a special meaning in the Budyko framework, the other boundaries for the bins do not necessarily need to be chosen because they are round numbers or evenly spaced. If instead you had chosen boundaries on each axis that contained approximately equal numbers of grid cells, you may arrive at a more robust and clear result with fewer dependencies on varying sample sizes. But I would suggest keeping a boundary for aridity at 1.0.

C20: Thank you for the suggestion. We have tried to use different aridity and temperature classes to obtain approximately equal numbers of grid cells in each bin as can be seen from the heatmaps below. We will update the aridity binning as shown in the first-row plots in order to have more similar number of grid cells in each box, and apply this binning for all heatmap visuals we use in the manuscript. The second row shows the previous version that we had in our manuscript.





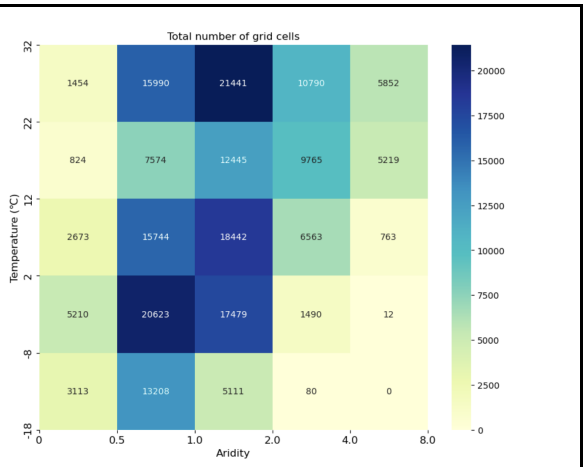


Figure A3 Total number of grid cells in each temperature-aridity category

17. L181-192: I appreciate this paragraph. If you are interested in pursuing this further, you might consider using an approach based in information theory, which has the advantage of also being nonparametric. There are also ways to quantify nonlinearity and parameter interaction (see: https://doi.org/10.1002/2016WR020218, https://doi.org/10.1002/2016WR020216, https://doi.org/10.1029/2020WR028179).

C21: Thank you for this suggestion. We appreciate the reference to information theory and its advantages. However, at this stage, pursuing this further in that direction would extend beyond the current scope of our paper, but we acknowledge that information theory-based methods could be an interesting approach for future studies.

18. L190: I think the independence of different data sources could be looked upon as a strength, not a weakness, of this research. When patterns emerge across datasets with different algorithms, or not all from one model, it gives more credence to the results.

C22: Thank you for this comment. We agree that the use of independent data sources can indeed be viewed as a strength of our study. We will mention this also in the discussion section. Here's a revised version:

"Another limitation is the data quality of each driver variable. A lower signal-to-noise ratio for certain variables compared to others may affect the identification of analogues and related temperature anomalies, and consequently the estimated relevance of the variable. However, the use of independent data sources can also be considered a strength of our study. We observe consistent patterns across different datasets which enhances credibility to our results and align with the existing literature on land surface and atmospheric patterns."

19. L205: Replace "highlight" with "highlights".

C23: "This finding highlights that the land surface generally affects hot extremes at longer time scales, as opposed to the more immediate influence of atmospheric drivers."

20. Figure 5: The result is not compelling unless statistical significance of these differences between decades can be established. Fortunately, that is straightforward. A very robust test is a bootstrap approach where the 20 years are randomly split into 2 sets of 10 and

the “degree of relevance” calculation is repeated many times (say 1000 times; C(20,10) = 184756, so no problem with oversampling). Then find where the particular case of 2001-2010 versus 2011-2020 falls in the larger distribution… that is your p-value. Otherwise, we don’t know if the EF changes are meaningful.

C24: Thank you for suggesting a method for the statistical significance of figure 5. We will implement a bootstrapping analysis as suggested. See response B5 for details on the planned approach.

21. L245: Drop the word “wide” – it’s not very appropriate.

C25: We have revised the sentence to read: “This study provides a comprehensive analysis of the potential drivers of hot extremes, considering a selection of atmospheric and land surface variables.”

22. L246: You say “particularly at the 500 hPa level” but the results for other levels were never quantified, save for squinting at all the similar shades of color in Figure A1. A Table (A1, perhaps?) should be included with the complete quantifications (for all factors in each decade – as Figure A7 is also difficult to read).

C26: Thank you for the suggestion. We will include a table in Appendix A that provides a complete quantification of the variable percentages for all factors at each level and for each decade.

23. L267: Here you are talking about trends, but you do not use the word “trend”. It would be clearer if you did.

C27: We have revised the text to include the word 'trend' for clarity.
“Another interesting result of our study is the positive trend in the relevance of the land surface in general, evaporative fraction in particular, driving hot extremes during the study period. This is likely related to higher temperatures and precipitation variability, which enhance the role of evaporation in the surface water and energy balances.”

24. Figure A2: Please expand the acronym “SHAP”.

C28: We will expand the acronym 'SHAP' in the figure caption to improve clarity.
“Relative importance (Shapley Additive Explanations, SHAP values) of multiple factors to explain the spatial patterns of geopotential height and EF as main drivers for 1-day and 7-day hot extremes.”

25. Figure A7: I suggest for the bottom 2 panels, only color the grid cells where a change has occurred. Leave the unchanged cells blank.

C29: Thank you for the suggestion, we will revise the figure accordingly.

26. Figure A8: Presumably there is a bit of movement of some grid cells between bins from one decade to the next. Are you considering that here, or are the 2m temperature and aridity still based on the 20-year climatology? Also, here again, a bootstrap statistical test can tell which bins have significant changes (or perhaps use color to indicate p-value).

C30: We are not considering decadal changes in aridity and mean temperature, but use the 20-year averages to create the bins. We will test to do the binning for each of the two 10-year periods to see if there are noteworthy changes in the number of grid cells per bin.