Comment: “Dear authors, I would like to thank the careful consideration of the comments raised by the reviewer. I have conducted an editorial review of the paper following their recommendation, and while I agree with the reviewers that original comments are well addressed, there are some improvements that need to be made in the clarity and writing style. Also, in reviewing the paper I was somewhat confused by certain aspects which can I am sure be resolved through comments as well as inclusion in the supplementary material. Please see the comments below. Most are editorial, though some are on the methods used.”

Reply: Dear Editor, we would like to thank you for your editorial review of the manuscript. We appreciate your recommendations and suggestions on how to improve it. Please find our replies to your comments below.

Comment: “Lines 30-33: I would remove the word "usually". For meteorological drought it may be better replaced with "often". For other drought types this would seem superfluous”

Reply: We have now removed this term and replaced it as suggested (lines 30-34).

Comment: “Line 39-41: Consider replacing the word "might" with "may". In UK English the first includes some doubt if this will happen, which is not what is intended”

Reply: We have replaced the word (line 41).

Comment: “Line 41: "currently accounts"”

Reply: We have made this change (line 41).

Comment: “Line 47: I am not so clear what is meant with "address the water use priorities", and think this can be best dropped, refering just to the need to reinforce water management”

Reply: We have removed this statement (line 47).

Comment: “Line 56: on research and practice of drought indicators”

Reply: We have made this change (line 56).

Comment: “Line 75-90: I appreciate how this is formulated, but it is also somewhat unorthodox. Consider converting this into a paragraph that outlines the objectives of the research”

Reply: We have now restructured these lines into a paragraph that clearly states the objective of the study (lines 76-89).

Comment: “Line 89: Would it be more appropriate to say "ways to take incomplete reports into account"”

Reply: We agree and have reworded this (line 87).

Comment: “Line 91: it would appear to me that the Methods section contains more than what is suggested here. Please complete”

Reply: We now say: “in Sect. 2 we introduce the study area, the drought indicators, climate indices and vulnerability factors used and their data sets. Also, how we deal with incomplete impact data and the methods for the data analysis.” (See lines 90-92).

Comment: “Line 100: what is meant by a high "concentration". Is this a high intensity of rainfall?”

Reply: to clarify this we have replaced the sentence in line 101 with “high interannual variability, with a high amount of rainfall occurring during relatively few days”.

Comment: “Line 101: rain-less periods are normally referred to as dry periods”

Reply: we have changed this (line 102).

Comment: “Line 102: I am not sure that aridity is a problem from the climate side. It is a characteristic of the climate. Just say "An arid climate, meaning....)”

Reply: we have changed this (line 103).

Comment: “Line 107: time --> temporal”

Reply: we have changed this (line 108).

Comment: “Line 107: The sentence starting with "Furthermore" does not make sense, grammatically. It is incomplete.”

Reply: we have replaced the word with “also” (line 108).

Comment: “Line 112: "evacuation plans"”

Reply: we have changed this (line 113).

Comment: “Line 124: consider replacing the first "studying" with "characterising" to avoid the repetition”

Reply: we have replaced it (line 125).

Comment: “Line 133: droughts do not limit the capacity of the infrastructure itself (as that does not change due to drought). It may be better to change to "shown to reach the limits of the capacity"....”

Reply: we have changed this (line 134).

Comment: “Line 138: "investigate their performance against the other types". Does not make sense - "compare their performance"?”

Reply: we have replaced the word (139).

Comment: "Line 148: The description of SPEI is not as clear as it should be. The word also suggests that SPI is based on the water balance, which it is not. SPEI is to my mind the balance between precipitation and evaporative demand. I think it is important to be clear that it is the evaporative demand, and not the actual evaporation. I agree it is better explained in later sections, but here it is confusing."

Reply: we have now replaced the first sentence to not suggest that the SPI is based on water balance. However, because the SPEI is the difference between precipitation and potential evapotranspiration (which is not the same as the actual evaporation). We continue to use the term 'potential evapotranspiration' instead of 'evaporative demand' (see Vicente-Serrano et al. (2010) for this definition of SPEI).

In lines 150-159, we now better explain the SPEI by saying: "The SPEI is a similar index to the SPI, but instead of being computed with precipitation values only, it is based on climatic water balance. The climatic water balance is a monthly difference between precipitation and potential evapotranspiration (PET) at different time scales. This provides a measure of the accumulated water surplus or deficit. We used the approach of Vicente-Serrano et al. (2014b) to calculate the PET; this is a simple approach that only requires data for monthly-mean temperature and uses the Thornthwaite equation (Thornthwaite, 1948). To obtain the final index, the same procedure as for the SPI was followed, however, a log-logistic probability distribution was used to model the precipitation–PET values. We calculated the SPEI also with the "SPEI" R package. This index accounts for the effects of temperature variability on drought. The advantages of this index, especially under global warming conditions, are that it identifies increased drought severity when the water demand is higher as a result of increased evapotranspiration. In addition, its multi-scalar nature allows its use for drought analysis and monitoring (Vicente-Serrano et al., 2010)."

Comment: "Line 163: It may need to be clarified why there are four layers. This is stated as being obvious but it is not clear where this comes from. I assume it is due to the ERA5 datasets used which indeed has four layers (again stated later)."

Reply: we have rewritten the paragraph slightly to clarify this (lines 163-164).

Comment: "Line 206: "drought impact models""

Reply: we have changed this (line 212).

Comment: "Line 243: These problems are not always clear. For example, for the 2nd, it is not clear what the issue is; but I assume that it is that the start and end month is not indicated. So please be more explicit on what the problem encountered is. In this case it would be that start and end year is indicated, but there is not indication of the month(s)"

Reply: we are now more explicit when explaining the problems. We have clarified this for all the cases (lines 249-253).

Comment: “Line 264-265: I assume that the reference to the S region containing 8 categories is to the least censoring approach (CM4). Unless interpreting this incorreclty, I would observe, however, that other regions also have more than 3 categories (NE, CE, E). So while the intent of the sentence is fine, the statement somewhat confusing”

Reply: the reference to the S region containing 8 categories is actually for CM1, not CM4. It is hard to see unless one zooms in. We now indicate we are referring to this CM in the sentence to avoid confusion (line 274).

Comment: “Line 271: add in brackets what the Entidad Estatal de Seguros Agrarios is for those not familiar with Spanish (e.g. National agricultural insurance agency - but check the official translation they may use)”

Reply: we now have added this in line 281. We could not find an alternative official translation.

Comment: “Comment: “Figure 3: The bars indicating the impact occurrences are not included in the legend (maybe clearer to include in the caption). Confusingly, the colour of these bars is changing. I would suggest to ensure the z-order of these is such that they are on top and non-transparent. I would also suggest to change the name of the y-axis to the Number of drought impact occurrences. Also the caption should be the "Number of monthly DIOs".”

Reply: The bars in Fig. 2 and 3 are now non-transparent, their y-axis label is changed to "Number of monthly DIOs" and their captions now link the bars to the impact occurrences.

Comment: “Line 309-313: Include variables in an equation, as this will give them a distinct font making clear that these are variables (in LaTeX you could enclose in $$)”

Reply: we have changed this (lines 319-321).

Comment: “Line 328: It is not so clear what the outputs are to which these four thresholds are applied. In my interpretation this is the normalised number of impact events. Would the predictand then not also be expected to be between 0 and 1? And why are there four thresholds, some of which are > 1. Please clarify.”

Reply: the values of the thresholds are not normalised. When converting regression model predictions to binary classes, we first normalised the thresholds, and then depending on whether the prediction was greater or smaller than the normalised thresholds, we converted to “no impact” or “impact” accordingly. The thresholds were normalised by dividing them by the total number of DIOs for each region (total for the whole time series, not just the training set). Some thresholds are > 1 because they are not normalised. If we use non-normalised thresholds to describe these, we can get a better feeling of the magnitude of the thresholds in the units of impacts. We now clarify this in the manuscript, see lines 339-340.

Comment: “Line 343-350: Correlations are discussed here for the agricultural indicator. It is curious to note that for all four layers, the correlation in the NE region is positive and

significant at all depths, while for other regions it is negative. There are some other exceptions in the E and CE regions. Is there any explanation for this anomalous pattern?"

Reply: As previously discussed in line 345, the NE and CE regions are the least populated regions in Spain, which could explain why we find the weakest negative correlations - there are less impacts reported because there is less exposure. This argument may also be supported by the fact that the NE region usually shows the weakest correlations with the meteorological and hydrological indices (see Fig. 4), indicating a weaker relationship between these types of drought and drought impacts.

The fact that positive and significant correlations happen generally at longer timescales (36-48 months, except for the deepest layer) and not at the same aggregation scales than the other regions' strongest negatives, could indicate that we are not just seeing an opposite relationship (high soil moisture = drought impacts), but that the apparent link has (1) a different physical explanation, (2) is simply non-existent but results due to a lack of data and/or coincidences between peaks and troughs of both time series, and/or (3) is due to your later comment on longer aggregation periods showing better results.

However, a closer look to Fig. 3 reveals that the NE has DIOs in July-October 2009, a period where there are no DIOs in other regions. This could also explain or be contributing to the positive correlations we are discussing (i.e. soil moisture increases and reports increase, and in the other regions the same effect is not visible since there are no impacts reported during this period).

We would like to introduce an example of how the effects of drought can be very long-lasting and show the known 'creeping phenomenon' to explain why these July-October 2009 DIOs may appear and help illustrate our next point:

In December 2009, several years after the severe drought event suffered in 2005/2006, it was reported that river basins continued to suffer marked impacts for water supply, both for the public and for farmlands and livestock farming. As a result, it was necessary to adopt urgent measures to mitigate the effects of the severe drought event.

The water year 2004/2005 started with good water storage levels, however, it had very low precipitation levels with respect to the historical mean. The next water year 2005/2006 had higher precipitation levels than the previous but still less than the climatological mean. This mean there was not sufficient precipitation to recover from the effects of the previous year, and at the end of the year, the water storage levels were even lower than for the previous year. The precipitation levels in the next water year 2016/2017 were higher than the climatological mean in spring in some regions but at the end of the year, the water storage levels were at the same concerning levels. During the following year, 2007/2008, average precipitation was still low (mostly in the S) so the water deficit from the previous years was not solved. In the following year, 2008/2009 mean precipitation levels were still below climatology. Water supply met the demands of that year (although precautionary measures were taken), but the irrigation campaign of that year was still conducted with difficulties. The following 2009/2010 year started with relatively low reservoir storage levels. This led the government to impose effective measures (Boletín Oficial del Estado, 2009).

This shows how a drought event that started in 2005 was still having long-lasting effects in 2009/2010. It is an example of how drought can be very long lasting (may explain the reported DIOs in the NE region in July-October 2009) and it can explain why indicators (especially ones that respond to changes in precipitation fastly) and impacts may sometimes not be negatively correlated. A region can still be suffering from drought impacts during a period where precipitation levels are above the climatological mean. This combined with inconsistent and incomplete impact reporting can then result in positive correlations in our data. It might give the impression that impacts suddenly occur during a period of higher precipitation/soil moisture whereas in reality, the impacts are not occuring due to this increase but due to the long-lasting effects and "creeping-phenomenon" of drought. This case example also reinforces the conclusion of why long aggregation periods could be better at modelling drought impacts.

We now give a summarised explanation to this anomalous pattern in lines 554-561.

Comment: "Figure 4: It would help to clarify better what is implied with aggregation period. In my understanding this means the 1-month aggregation period you are exploring correlations with is in fact SPI-1, and for the 48 months this would be SPI-48 (same for other indicators). I think it would help the reader if this is added to methods section, as it is not 100% clear if the aggregation is done prior to the fitting of either the gamma or log-logistic distributions and transforming to the normal distribution, or after. However, it is not then clear how this is done for the climate indices (NAO, etc). Are these then averaged over the aggregation periods indicated as a moving window (ie a post averaging)?"

Reply: we have now added: "All of the indices were aggregated over different time scales. The aggregation of the SPI, SPEI, SSFI and SRSI was done prior to fitting them to a distribution and transforming to the normal distribution. This means the data for the current month and past X months was used to compute the value for a given month The aggregation period is hereafter labelled with '-X' (e.g. SPI-X). Similarly, for the SWSI, the aggregation was done prior to computing the empirical distribution." (See lines 168-171). And "All of the climate indices were aggregated by computing a moving average over X months" in lines 179-180.

Comment: "Figure 5: This figure could be included in the supplementary material as it is only used to support the statement on line 356. This will also reduce article length (and cost)."

Reply: we have made this change, it is now Fig. S2.

Comment: "Line 389: I am somewhat confused at how it is possible to obtain such a small value of RMSE. If the predictand is the normalised number of impacts, and there are e.g. 3 impacts categories, then the possible values of the predictant would be 0, 1/3, 2/3 or 1. As the R2 is not perfect, a mismatch of 1 predicted impact would result in an error of at least 1/3. Would this not result in a much larger RMSE? It may be useful to explain (perhaps in supplementary material) how these statistics are calculated."

Reply: The RMSE values are normalised, if they are multiplied by the total number of DIOs in a region, then one can obtain the value in units of DIOs. This multiplication yields, for

CM1, RMSE values of: 0.1, 0.5, 0.1, 0.5, 0.3 and 0.5 for the regions NW, NE, MA, CE, E and S respectively. The models tended to underpredict the number of DIOs and this explains why these values are small. The time series of DIOs does not contain a lot of DIOs compared to the (larger) number of time steps with no DIOs. The RMSE is small because the model is good at not predicting DIOs when there are none (high specificity) and because there were more actual DIOs than those predicted; this occurs for large parts of the time series. The models also had a very high precision, meaning that the predictions of impact occurrences were usually correct. We now explain this in lines 403-409.

It is not possible to determine the values of the predictand as you comment, since the predictand is an average of all the tree predictions made by all the decision trees in the random forest and not limited to certain values. The predictand is only categorical for the classification models, since the output is binary and is determined by the most popular class voted by all decision trees.

We now explain how all of the performance metrics are calculated in the cross-validation analysis in the supplementary material (S1) to clarify this.

Comment: "Figure 8: In line with the previous comment, I do not understand how the AUC can be 1.0 (such as for the E region under CM1), while the recall is < 1. A recall of < 1 in my understanding means that there are missed events (ie impacts did occur but were not predicted).A precision of 1 indicates there are no false alarms (as also indicated in the text in line 402). However, as there is at least 1 missed event, the probability of detection (number of events predicted / total number of events), which is equivalent to recall must also be < 1, and thus the AUC cannot be 1. Also, the uncertainty of the estimate of AUC is > 1, but the indicator can only range from 0-1. Same holds for recall and precision. Perhaps I am missing something, but the results are somehow confusing and require some clarification."

Reply: We tuned the models by selecting the *mtry* parameter that yielded the best model in terms of the performance metric being assessed. For example, to compute the precision (or another metric) of a model, the *mtry* parameter value that yielded the predictions with highest values of precision (or another metric) was chosen. A new model was run each time (to compute each performance metric) which meant that different *mtry* parameters were used. Because of this, the predictors randomly sampled at each split (*mtry*) were different. This means that we cannot compare the different metrics for a region as you do in your comment. The predictions made by each model for each testing part of the data will always vary slightly (i.e., the models used to produce precision plots are not the same as the ones used to obtain the recall or other plots since the former uses the precision metric to tune the model and the latter uses recall instead). One would be able to compare the results as you are doing only if the models were all tuned using the same performance metric, meaning that they would output the same predictions for each split of the data. The splits of the data into folds were kept constant for all runs.

We now explain this in lines 428-430 and together with the previous comment, we explain how these statistics are calculated in the supplementary material (S1).

Comment: "Line 447: The thee subsequent sentences start with "When" please try and introduce a little more variation."

Reply: we have made this change (lines 510-514).

Comment: "Line 456: This conclusion/finding is somewhat trivial as the counting of an impact in all regions is inherent to the design of CM4."

Reply: We have removed this statement from the manuscript (line 473)

Comment: "Line 516: "both analyses we found overall". Try also to use another word for the second "found" in the sentence."

Reply: we have made this change (line 532).

Comment: "Discussion and Conclusion section: Whilst I think the general discussion and conclusions are supported by the study (as also confirmed by the referees), I am somewhat concerned/confused about some of the statistics presented and think these need additional explanation. I am not sure based on these how firm some of the conclusions are."

Reply: In the following replies, we address this general comment.

Comment: "Reviewing the correlations found in Table 3a and 3b, it appears that these are, with a few exceptions, very close in value (e.g. for CM1 in NW region the range of correlation values is 0.04, which is small). This is also clear when comparing to CM2, where a small change is introduced, which then completely reverses the order of predictors in terms of correlation."

Reply: We do not seem to agree that the values are completely reversed. A comparison of CM1 and CM2 correlations, when looking at drought indices (first two columns) shows similar results - the most differences we see are in the NW where the soil moisture index has higher correlations (range of 0.01-0.04) in CM2 than in CM1 (ignoring the positive correlation since Fig.4 shows that it is not actually relevant). We still see that the SPI shows similar correlation values (0.06) in both CMs. The differences for the climate indices are slightly higher but usually the same predictors appear in both counting methods We now briefly explain this in lines 373-375.

Comment: "There is also a tendency, as reported, for longer aggregation periods to show better results. The discussion suggests that this is due to the propagation of drought through reservoir and groundwater systems. However, the discussion could also explore if this is simply an attribute of the methods used. There are not that many events, and as figure 3 shows impacts are often clustered (may be good to include a similar figure as fig 3 for CM3 and CM4 in supplementary material). Increasing aggregation periods tends to smoothen out the curves as well as "extend" periods that are considered as below normal as well as those considered as above normal. To my mind this would also mean that there is a higher chance of hitting the sweet spot for longer aggrgation periods, but this need not be a causal link. The results of in particular the regression and forest tree models show there is a tendency to be high on precision, but lower on recall - which would confirm this suggestion. Same holds for the selection of the thresholds, which obviously increase recall as these are set to lower values (but may induce false positives). I think these aspects need to be discussed in a little more depth."

Reply: We have added “It is important to note that considering that there are not many drought events within the study period and that impacts are often clustered due to the nature of drought events, the observed tendency for longer aggregation periods to show stronger links to impacts can also be attributed to the methods used. This could be due to (or partly) aggregated indices being more smoothed out and consequently extending the periods that are above and below the normal” (lines 571-574).

We have now added a figure similar to Fig. 3 but for CM3 and CM4 in the supplementary material (Fig. S1).

Comment: “The point raised by the non-importance of drought indicators in all regions except MA when vulnerability factors are included (line 466) is somewhat indicative of the sensitity of results and thus of conclusions; particulalrly as I assume most these factors are taken as being static. I would assume that MA as a largely urban area with a more services based economy shows different behaviour as many of the factors introduced are simply not relevant.”

Reply: We would like to highlight that the conclusion about the non-importance of drought indicators in nearly all regions when vulnerability factors are included is relevant for the vulnerability analysis but (most probably) not for the main analysis. The analysis that included the vulnerability factors was limited by the availability of vulnerability data. The period studied missed the first two drought events, hence, the robustness of the models built and their results is questionable, as previously discussed in line 464. The non-importance of the drought indices can therefore not be simply assumed for our main results since the models built with the drought indices and vulnerability factors were not as robust. We now emphasise this in lines 486-488 and have added that the “models are built with data from 2000-2012” in the caption of Fig. 11.

Comment: “I have also been wondering about the comment that the lowest soil layer has the strongest correlation at the shortest time scale. I cannot see a logical causal explanation to this as this layer is typically below the root zone of most crops. What I can imagine is that as it would be expected to have the lowest dynamic, there is already a natural aggregation, thus matching the higher layers with longer time scales. Longer timescales for an indicator that is already aggregated would then tend towards an average (with no anomalies), and thus lower correlation.”

Reply: We agree, and have removed the comment and explanation about the lowest soil layer having the strongest correlations at shorter time scales. We now include your suggested explanation in lines 551-554: “The agricultural index showed more significant and negative correlations in the two shallowest soil layers. A weaker link between lower layer soil moisture and drought impacts can be explained by the fact that the soil moisture content of these layers, which are usually below the root zone of most crops, has a slower and more aggregated behaviour. Aggregating these indices then creates a more averaged time series with less anomalies, which then leads to lower correlations.”

Comment: “I would recommend the authors to look carefully at the manuscript and some of these comments. Some may stem from confusion and simple clarification could help.”

Reply: Hopefully with our replies to your comments and changes to the manuscript, we have addressed your concerns and clarified certain aspects.

Comment: “I think it would be very useful to also include results datasets to support the paper. Currently the source data and packages used are referenced (which is appreciated), but interpretation of the results is not easy. The results of the various predictor models could be provided/explained to help interpret the resutls presented in the paper.”

Reply: We now have included the results datasets of the cross-validation analysis (line 622-623). It includes a short text file explaining how to tune RF models, to reproduce our results. We hope that this, together with the explanation of how the cross-validation analysis was performed (supplementary material) clarifies the results.

We thank the Editor again for his detailed and constructive comments, we believe they have helped improve the quality of our manuscript.

References

Boletín Oficial del Estado: Real Decreto 14/2009, de 5 de diciembre, por el que se adoptan medidas urgentes para paliar los efectos producidos por la sequía en determinadas cuencas hidrográficas, https://www.boe.es/boe/dias/2009/12/05/pdfs/BOE-A-2009-19563.pdf, 2009.

Vicente-Serrano, S. M., Beguería, S., and López-Moreno, J. I.: A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index, J. Clim., 23, 1696–1718, https://doi.org/10.1175/2009JCLI2909.1, 2010.