We would like to thank both reviewers for the overall positive evaluation of our work, and the constructive feedback. In the following, we respond to all raised comments individually. The responses are numbered, with the first digit indicating the reviewer, and the following digits indicating the order of the comments. We are confident that our thoroughly revised manuscript now addresses all raised points and hope for it to be considered worthy of publication in NHESS.

Line numbers refer to the revised manuscript (which is not uploaded yet, and line numbers might change)

2.	**Reviewer comment**
	In this study, the authors used a multi-index approach (exposure, risk, vulnerability) to model the impacts of drought events on agricultural systems in the German federal state of Brandenburg, considering the LST/NDVI ratio as the response variable. The scientific approach used is valid. It reflects the multifactorial complexity of the implications of drought for the productivity of the region's farming systems.
	Author's response
	Thank you for this positive overall evaluation. For the sake of completeness, although the reviewer is certainly aware of this, we might add here that we even used 2 different response variables: the LST/NDVI ratio (per field) and relative yield gaps (per county).
	Changes in the manuscript

2.01	**Reviewer comment**
	« Empirical associations to the impact indicators on both spatial levels are compared. Non-linear models explain up to about 60% variance in the yield gap data, with lumped models for all crops being more stable than models for individual crops». This is imprecise, you must specify the names of the nonlinear models as well as for the grouped models. It is also important to include in the abstract the performance statistics of the models used.
	Author's response
	While we agree with the reviewer that model names and numbers can be important in many cases, we have to find a compromise here because the Abstract is limited to 200 words, and we would like the key messages in the abstract to be generally understandable by interested readers with less technical focus. In response to this request, we specified the model type from "non-linear" to "XGBoost" and added the precise score of the best model on county level "(R^2 = 0.62)". The general statement about the stability of the models refers to the figure 13 in the manuscript, where it is shown that distributions from repeated training with samples from all categories have a higher mean and lower variance than models repeatedly trained on thematic subsets such as individual crops or years. There is no single number to report, though, rather it is a condensed finding that we attempt to convey in language. We rephrased this part to make the findings clearer, and hope that the reviewer may agree with this attempt, or otherwise we are open for suggestions on how to further improve the abstract.
	Changes in the manuscript
	L21 now reads: "XGBoost models explain up to about 60% variance in the yield gap data (best R^2 = 0.62). Model performance is more stable for the drought years, and when using all crops for training rather than individual crops."

2.02	**Reviewer comment**
	"Rye is found less vulnerable than wheat, despite growing on poorer soils". The fact that rye grows on poorer soils is a proof that it is more resilient and less vulnerable than wheat, so I do not see why the conjunction of subordination although?
	Author's response
	This is a very good point. In fact, we find that rye on poor soils still is less affected by drought than wheat on good soil. It is one thing to assume one crop to be more or less vulnerable, or to find this in laboratory experiments, and another thing to substantiate this empirically under real world conditions. We added a sentence for clarification.
	Changes in the manuscript
	L23: "Rye is empirically found less vulnerable to drought than wheat, even on poorer soils." L518: "While this already indicates that rye tolerates harsher conditions, we find empirically that rye on poor soil is still more robust under drought conditions in the region than wheat on good soil – based on both impact datasets."

2.03	**Reviewer comment**
	In introduction, « This has implications for modelling and Monitoring ». You mean implications in the modelling and monitoring of agricultural drought. If so, the sentence should be completed.
	Author's response
	Thank you for this observation. We completed the sentence accordingly
	Changes in the manuscript
	The line now reads: "This has implications for modelling and monitoring of agricultural drought"

2.04	**Reviewer comment**
	Overall, the introduction is well written and argued. However, the application of artificial intelligence models in modelling drought impacts, risk, and vulnerability has been limited. It is worth adding a paragraph on the advantages and limitations of intelligence models in modelling the impacts of drought given that in your methodology you have used the extreme gradient boosting algorithm (XGBoost).
	Author's response
	Thank you. We revised the paragraph on methods in the introduction and included more discussion on advantages and limitations of AI models for drought impact prediction. Many of the references in the paragraph on methods use algorithms from the AI domain, especially Kondylatos et al. 2022, Peichl et al. 2021, Merz et al., 2013, Brill et al., 2020, Sodoge et al., 2023, Tanguy et al., 2023, among others. Some parts in our Results & Discussion and Conclusion sections already addressed opportunities and limitations of AI methods, but in response to this reviewer request we made it more prominent in the revised manuscript.

	Changes in the manuscript
	L72: “from the field of (explainable) artificial intelligence (AI and XAI, respectively)” L82: “The application of AI methods in particular has led to considerable advances on the side on drought hazard monitoring and forecasting in recent years (Prodhan et al., 2022; Kowalski et al., 2023; Zhang et al., 2024). While these methods are very promising, they do rely on the availability of (big) data covering the processes of interest. On the side of vulnerability and impact-relevant factors, a key bottleneck of such data-driven studies is the availability of impact data.” L548: “We chose the algorithm XGBoost, which, compared to Random Forest, limits the amount of variability between the individual decision trees. This is assumed to avoid erratic behavior, but on the other hand could also limit the potential damaging processes discovered by the models. For the models on county level, predictive features were derived by computing the relative area above/below evenly-spaced thresholds. An alternative here would be to use quantiles, or to automate the feature engineering by deep learning algorithms. Stronger AI methods, not only in the regression but also in the feature learning step (i.e. deep learning), could improve the predictive skill.” L577 now reads: “Data-driven techniques from the AI domain can capture complex interactions in human-environments such as agriculture. SHAP plots uncover which factors drive the prediction of impact indicators in the models. This does not necessarily relate to causal effects in nature, though. We thus suggest to cross-check results obtained from different model setups, different regression targets, and ideally also different algorithms.”

2.05	**Reviewer comment**
	Line 250 «To retain as much information about the hazard distributions, we computed the relative affected area (non-)exceeding specified thresholds (in regular intervals of 0.5 for SPEI, 0.25 for LST/NDVI-anom., 0.05 for SMI, 5 for SMI-Total, and using the LBG class limits for AZL). A total of 68 features were created this way on county level». On what criterion were these thresholds considered? This deserves to be clarified. The different classification thresholds for these indices and their meanings should be provided in a table in the methodology section.
	Author’s response
	True, this is a good point to further detail in the methods and also mention in the limitations of our study. The thresholds were chosen in regular intervals, which is a form of manual feature engineering. An alternative would be to use quantiles (less intuitive), or do automate feature engineering altogether via deep learning techniques. We added this in the discussion on limitations and included a table as requested. If the overall manuscript becomes too long, we could also shift this table to the Appendix.
	Changes in the manuscript
	Table 2. Intervals for thresholds

Indicator category	Interval for thresholds (exact values)
SPEI	0.5 (-4*., -3.5*, -3*, -2.5, -2, -1.5, -1, -0.5, 0)
SMI	0.05 (0, 0.05, 0.10, 0.15)
SMI-Total	5 (0, 5, 10, 15, 20, 25, 30, 35)
LST/NDVI-anom	0.25 (0, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50)
AZL	LBGs (23, 29, 36, 46)

*only for SPEI-Magnitude

L549: "For the models on county level, predictive features were derived by computing the relative area above/below evenly-spaced thresholds. An alternative here would be to use quantiles, or to automate the feature engineering by deep learning algorithms."

2.06 **Reviewer comment**

The principle of the calculation of the LST/NDVI anomaly has not been sufficiently described. There should be a separate section to better describe and justify the choice of this anomaly to represent the impacts of drought when there are various other anomalies or indices that can better reflect the impacts of drought on agricultural systems. In this sense, the normalization indicated in Table 1 concerns only the LST values and/or the LST/NDVI values. If so, considering the max and min values or mean and standard deviation (SD)?

Author's response

It is certainly true that other indicators could be used for observing drought impacts, and we pointed this out once more in the revised manuscript by adding additional references on alternative metrics.

However, we are convinced that the provided literature references contain sufficient justification for using LST/NDVI. L91 in the preprint: "The ratio between LST and NDVI is a well-established observable indicator for that purpose (McVicar and Bierwirth, 2001; Karnieli et al., 2010; Crocetti et al., 2020). Mid growing season is generally regarded as the most decisive time of observation (Ghazaryan et al., 2020)."

In line with the reviewer comment, one of the conclusions of our paper is that better remote sensing indicators should be explored. L610 in the preprint: "Monitoring and impact-based forecasting are needed to prepare for future hazards, which can hardly be mitigated. Stronger remote sensing indicators on drought impacts seem necessary in that context."

The anomaly was calculated as difference between the field-level values and the area-weighted mean for a specific crop, so that values for different crops can be better compared. We included a formula in the manuscript as requested. In addition, we would like to mention once more that the exact procedure of our study is documented in the provided programming scripts, which are publicly accessible via Github

We would also like to highlight that we used 2 different impact indicators and compared them – the other one being based on reported yields. The strongest point of our study might be this comparison, which implicitly addresses potential shortcomings of individual impact indicators and insights derived thereof.

	Changes in the manuscript
	Introduction: "While there are various potential indicators for mapping drought impacts on crops, the ratio between LST and NDVI is a particularly well-established observable metric for that purpose (McVicar and Bierwirth, 2001; Karnieli et al., 2010; Crocetti et al., 2020)." New subsections at L195: "2.4.1 LST/NDVI Anomaly", and L212: "2.4.2 Empirical yield gaps" L207: " $$LST/NDVI_{anom,f,y} = \frac{LST/NDVI_{c,f,y} - \overline{LST/NDVI_c}}{\overline{LST/NDVI_c}}$$ (Eq.2) Where $I_{anom,\ f,\ y}$ is the $\overline{I_c}$ is the area-weighted mean for a given crop across all years, and the subscripts c, f, and y denote crop, field, and year, respectively. " L539: "The choice of impact variables, and preprocessing thereof, might introduce biases. LST/NDVI anomaly is a commonly used indicator for drought-related crop health, but others are possible, such as the radar vegetation index (Kim et al., 2012), hyperspectral metrics (Dao et al., 2021), fractional cover time series (Kowalski et al., 2023), or multimodal techniques (Karmakar et al., 2024)." L577: "Stronger remote sensing indicators on drought impacts, beyond LST/NDVI, seem necessary as well"

2.07	**Reviewer comment**
	Ligne 255-260 « In 2013 and 2014 the SMI-Total is close to 0, observed vegetation health is at its maximum (i.e. negative LST/NDVI-anom.), essentially no impact-related statements…..» Similarly, to better assess the consistency of these statements, the formula and principle of the calculation of the IMS and IMS-Total must be clearly described in the methodological section with the different classification thresholds.
	Author's response
	Thank you for pointing this out. We recognize that our description of SMI-based features was not as clear as it should be. We added the formula for the soil drought intensity and a description. More details of the calculation are described in the referenced article by Boeing et al. (2022). It might be slightly confusing that we use the SPEI absolute values, and the SMI-derived drought intensity, as well as the SMI-derived drought magnitude for the total soil, but still refer to the values as "SMI" rather than "SMI-based drought intensity". We are open for renaming the features (for example to "SDI" or "SMDI"), if the reviewers consider this more understandable. Please note that normalization does not have any effect on the XGBoost models, as they operate with relative differences rather than absolute values.
	Changes in the manuscript
	L179 now reads: "Identical to the SPEI data, we use monthly values and a growing season aggregation of drought intensity derived from the soil moisture index (SMI) for the top soil (25 cm), again from March to July. To add some information on slower long-term drought processes (i.e. accumulation and lag time), we further include the annual drought magnitude for the total soil (up to 1.8 m depth), which

	is temporally aggregated from April to October (SMI-Total). SMI drought intensity represents the integrated area below the 20th percentile of the soil moisture index for a given time (and area). The general formula, as presented in Boeing et al. (2022), includes a potential normalization over the area of investigation (Eq. 1). In contrast to the drought intensity, the drought magnitude is not normalized (i.e. shift of absolute values but same relative order). However, in this study we used the individual raster cell values, which implies that no area normalization is performed either way. $$SMI = \frac{1}{d \cdot A} \sum_{t_0}^{t_1} \int_A [\tau - SMI^*_i(t)]_+$$ where τ is the drought threshold, SMI* is the raw soil moisture index, and d and A refer to the duration and area of potential aggregation, respectively. A value of 0 for all SMI-based features thus means, that none of the values were below drought threshold τ. We use τ = 0.2 (20th percentile), which is a common value for drought analysis adopted in the literature (e.g. US drought monitor, Svoboda et al, 2002). For more details, the interested reader is referred to Boeing et al. (2022).“

2.08	**Reviewer comment**
	In Table 1, you mentioned that the monthly SPEI used has a resolution of 10 km and the source is the reference Zhang et al. (2024). However, in this reference, the SPEI used has a 1 km resolution. It is a bit ambiguous. Has the SPEI been calculated? or was the same database from the Zhang et al. (2024) study used? If this is the case, the spatial resolution of 10 km should be rectified because in the source reference mentioned it is rather 1 km that is mentioned.
	Author’s response
	Thank you for this observation. We apologize for the confusion. We used the same method and code to recalculate SPEI as in Zhang et al. (2024), including data transformation and quality check, but the data are actually on 10 km, based on the 0.1° spatial resolution E-OBS dataset by Cornes et al. (2018). We revised Line 165 in the main manuscript to make it clear for readers. We added the proper reference (Cornes et al., 2018) in “Data source and references” in Table 1 and the Reference section. Author HZ found the accuracy of this SPEI dataset to be higher than the one used in her previous publication (Zhang et al., 2024), and now also uses this in another ongoing study (Zhang et al., in review). The data can be viewed in the provided R-Shiny app.
	Changes in the manuscript
	L170 now reads: “Monthly values of SPEI-1 (one-month accumulation SPEI) used in this study are at a 10 km grid resolution from 2013 to 2022 based on the E-OBS dataset (Cornes et al., 2018). The calculation detail can be referenced in Zhang et al. (2024)." Table 1: “(Cornes et al. 2018)” References: “Cornes, R. C., et al. 2018. An ensemble version of the E-OBS temperature and 239 precipitation data sets. Journal of Geophysical Research: Atmospheres, 123(17), 9391-240 9409.”

2.09	**Reviewer comment**
	The algorithm used to calculate the Landsat LST was not explained in the methodology
	Author's response
	The data was taken from Google Earth Engine Landsat 8 L1T2. We included a reference to the procedure (Cook et al. 2014), although the data description from the USGS might be more exhaustive. The reference is in the Table 1. The respective code in online on GitHub. In response to this reviewer comment, we made the data reference more obvious in the manuscript. Data: https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1_L2#description Code: https://github.com/fabiobrill/brandenburg-drought-study/blob/main/preprocessing/gee_landsat8_dlschema.js
	Changes in the manuscript
	This dataset already includes processed LST (Cook et al., 2014). Table 1: "https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1_L2#description"