The reviewer's comments and the authors' responses are provided below, with the comments shown in black and the responses in red.

Reviewer #1

The manuscript "Towards impact-based early warning of drought: a generic framework for drought impact prediction in the UK" is well written and presents a relevant study. The authors evaluate multiple models for predicting drought impacts, carefully train and validate them with different lead times. Overall, the study is rigorous, the results are clearly presented, and the framework seems solid and well analyzed. I identified two shortcomings that should be revised.

>> We thank the reviewer for the positive assessment of the manuscript and appreciate the constructive feedback, which we addressed to further improve the work.

First, the discussion of multicollinearity is unsatisfactory. Readers need to trust the authors' assertion that multicollinearity is present but varies. Finally, this concern is brushed aside in line 605 with the sentence: "Although SPI and SPEI indices are highly correlated in the UK, the RF model is capable of managing this multicollinearity." However, multicollinearity is indeed a problem for Linear Regression and LASSOCV, which are compared in the first step before being outperformed. A more explicit quantification is necessary (e.g. showing a correlation matrix or variance inflation factor).

>> We agree that a more explicit quantification of multicollinearity would strengthen the discussion and improve transparency. In response to the reviewer's suggestion, we conducted additional analyses to demonstrate the extent of interdependence among the predictors used in the linear models. Specifically, we quantified multicollinearity for both the Linear Regression (LR) and LASSO models by computing Variance Inflation Factor (VIF) values for all predictors (*Supplementary Document* - Table S3), which revealed substantial multicollinearity among the drought indices for LR, with many VIF values exceeding 30–60. We also prepared a correlation matrix (*Supplementary Document* - Figure S3), confirming that numerous SPI, SPEI, and SSMI variables are highly correlated across time scales.

The manuscript is revised accordingly;

In the Method section we included these two new analyses and their aims in lines between 340-343:

"Beyond the performance metrics, we also assessed the multicollinearity of the predictors in the linear models by calculating the Variance Inflation Factor (VIF) and the Pearson correlations among the drought indicators."

And the discussion of these additional analyses is included in lines between 614-620.

"SPI and SPEI indices are highly correlated in the UK, and our analysis shows that this multicollinearity substantially affects the performance of LR and LASSOCV. The correlation matrix of predictors (Figure S3) revealed strong interdependence among drought indices across time scales, and the Variance Inflation Factor (VIF) values for the LR model (Table S3) frequently exceeded 30–60, indicating severe multicollinearity. LASSOCV also failed to mitigate this issue: the best penalty parameter selected via cross-validation was $\lambda = 0$, effectively reducing the model to an unregularized linear regression. These results showed that both LR and LASSOCV performed with lower accuracy relative to other ML approaches for the UK because they cannot reliably disentangle the effects of highly correlated predictors."

Second, the dataset (1970–2012) is quite outdated, even though more recent data (up to 2024) seems to be available, as noted in the data availability statement. This weakens the study's relevance and raises questions about whether it still represents state-of-the-art work. This is particularly disappointing since the introduction references more recent drought events (2018–2019: Turner et al., 2021; 2022: Barker et al., 2024) and sets expectations that are not met when entering the methods section. I suggest clearly stating the training period already in the abstract (e.g. "trained with data from 1970–2012") and including a discussion on whether the model would be capable of forecasting more recent, unprecedented droughts. Ideally, if possible, predictions for these years could be shown.

>> We appreciate the reviewer's concern regarding the dataset. The impact dataset used in this study is derived from the European Drought Impact Inventory (EDII), for which our team provided and curated the UK component. At present, this represents the most comprehensive and peer-reviewed source of drought impact data for the UK. Although preliminary data for more recent years exist, they have not yet been fully processed, validated, or published, and therefore were not used in the current study.

We now stated the study period (1970–2012) in the abstract to improve clarity, in the lines between 13-14.

"Here, we used data from the European Drought Impact Inventory (EDII, 1970-2012), and a wide range of meteorological and hydrological predictors, including the Standardized Precipitation Index (SPI), Standardized Precipitation-Evapotranspiration Index (SPEI), and soil moisture indices (SSMI), to develop a generalized forecasting framework for predicting drought impacts in the UK across multiple lead times."

Finally, following the suggestion of the second reviewer and to address concerns regarding potential confusion between EDII and EDID datasets, we removed the EDID data citation and reference from the manuscript, as these data were not used in the present study.

Minor comments

• Defining the upper tercile as "extremes" seems overstated, particularly since the threshold includes the upper 33%.

>> We agree that referring to the upper tercile as "extremes" may overstate the nature of the threshold, since it includes the top 33% of values rather than truly rare events. To address this, we retained the original naming convention from Shyrokaya et al. (2024) for consistency with the literature, but we now describe the terciles as minor, moderate, and significant, corresponding to the lower, middle, and upper terciles, respectively. Specifically, the lower tercile represents minor impacts, the middle tercile moderate impacts, and the upper tercile significant impacts. This terminology more accurately reflects the statistical distribution of the data while preserving the impact-based focus of our study.

We have also included these updated category names and the accompanying explanation in lines 310–317 as follows.

"After generating predictions for the monthly number of drought impact reports, we converted these predicted values into drought severity measures using the percentile methodology, which preserves the true scale of impacts while addressing biases present in reports across different regions (Shyrokaya et al., 2024). Predicted values were ranked within the historical distribution of predicted drought impact reports for each NUTS1 region, and a corresponding percentile rank

was assigned. The conversion used the tercile methodology, with the lower, middle, and upper terciles now referred to as minor (0–33%), moderate (33–66%), and significant impacts (66–100%), respectively; this updated terminology replaces the original names from Shyrokaya et al. (2024) to better reflect the statistical distribution of the data while maintaining an impact-based interpretation. "

• Several passages are overly long and difficult to follow due to heavy nominalization. Please streamline them for clarity or remove if they add no value. Examples: Line 60/ Line 142/Line 449

>> We reviewed the manuscript and revised the identified passages (Lines 60, 142, and 449) to improve clarity and reduce heavy nominalization.

Lines 60:

New sentence in Lines 59 >> "*In a recent review, Shyrokaya et al. (2023) surveyed efforts to develop drought impacts forecast. They found that most studies were academic, with only a few practical applications. The review also noted that, although operational systems remain limited, the past decade has seen substantial growth in the science supporting impact-focused DEWS and drought impact forecasts.*"

Lines 142 :

New sentence in Lines 142 >> *". Moreover, most early studies pre-processed the data by censoring certain time periods when selecting drought events, which limited their ability to assess model accuracy over longer timeframes and reduced their applicability for operational decision-making.*

Lines 449:

New caption in Lines 445 >> "*Observed and predicted drought severity categories (panels a, c, e, g) and likelihoods of occurrence (panels b, d, f, h) for selected UK NUTS1 regions (1970–2012). In all panels, the negative y-axis shows predicted values and the positive y-axis shows observed values. Grey shading in panels a–d indicates the temporally independent validation period (2007–2012), while panels e–h (South East and East of England) indicate spatially independent validation.*"

The reviewer's comments and the authors' responses are provided below, with the comments shown in black and the responses in red.

Reviewer #2

Dear Burak and colleagues,

apologies for the delay of this review and thank you for the opportunity to comment on this manuscript. I enjoyed reading about your study and provide my comments below. They are minor.

Regards
Kerstin
(Kerstin Stahl)

Summary

The manuscript reports a study that applies a range of advanced statistical models to the task of predicting drought impact occurrence under certain meteorological and soil moisture conditions in the UK. To do that it used training data from a text-based category-coded impact database and indices from operational monitoring or hydrometeorological conditions. The contribution is a valuable application that tests and compares various statistical modelling options that have not previously compared to that extent. The study also evaluates the potential operational use of this application, specifically prediction/forecasting of impacts with various lead times. The paper is well generally well written and makes an important contribution. A few aspects need some improvement to provide a more focused and consistent message and therefore the impact the paper deserves. They should be fairly easy to implement.

>> We thank the Kerstin Stahl for her positive assessment of the manuscript and appreciate the constructive feedback, which we will address to further improve the work.

Main comments

The title is quite long. In light of potentially misleading terminology, I suggest that 'impact-based' as well as "generic" might be removed from the title. But there may also be other solutions.

>> We thank the reviewer for this suggestion. After clarifying the terminology used in this study regarding the generic framework and impact-based forecasting, We have revised the title to:

"Toward early warning of drought impacts: a framework for predicting drought impacts in the UK"

My main comment relates to those two terms and the focus and consistent message of the paper. The two terms come with ambiguity and are used very differently in the literature. I think their precise use could be improved throughout the manuscript.

(1) "impact-based forecasting" is used interchangeably with "impact forecasting/prediction" in the title and text. While cited sources have used that term, some of the literature on climate impacts uses "impact-based-forecasting" differently, e.g. selecting ensemble members of a physical model based on impact information; the impact, however, is then not directly forecasted. Strictly speaking, "impact forecasting/prediction" might therefore be more correct. However, at least consistency and introduction/discussion might be improved on that.

>> We thank the reviewer for this important comment. To ensure clarity, consistency, and alignment with standard usage in the literature, we have replaced all instances of "impact-

based forecasting" in the title, headings, and text with "impact forecasting" or "impact prediction."

(2) "generic framework" suggests to me a standardized procedure of applying this in operation as indicated that it will be and/or a procedure that is transferrable to other data and regions. It speaks a bit against the rather detailed analytic comparison and analysis of several statistical models and the different forecasts that is the main aim and in fact in my opinion the main value and contribution of the study.

For consistency in the aims and main contribution made with the study, I strongly suggest to either tone down this 'generic framework aim' or explain in more detail what it is exactly in the end - perhaps including a flow chart or so. The methods generally have been applied previously, so what the general methodological 'developing' is, might also be clarified.

Figure 1 goes a bit into that direction but is not entitled "framework". So which part of it is the framework? And would that operational framework always use all model options? i.e. train all, but then apply/predict with the best? Or how is this transferred to the framework of application?

>> We thank the reviewer for this insightful comment. In the revised manuscript, we have clarified the term "generalized framework" (formerly "generic") in the title and throughout the text to better reflect its intended meaning. The generalized framework specifically refers to Stages 1 and 2 of our workflow (new Figure 1) Model Development and Evaluation & Comparison which define a standardized procedure for developing and evaluating models using UK-wide, lumped data. This approach is generalized in contrast to most previous studies, which develop region- or sector-specific models, and allows for systematic assessment across all regions while ensuring robustness and transferability.

Stage 3, Prediction & Forecasting, represents the operational application of this framework, in which only the best-performing model (selected during Stage 2) is applied for near-real-time prediction, multi-lead forecasting, and generation of gridded drought impact maps.

Finally, the transferability of the selected model is further demonstrated by application to Germany, chosen due to the availability of EDII datasets and prior studies on drought impacts.

These revisions ensure consistency between the study aim, the description of the framework, and its operational application, while clarifying that the novelty of our work lies in the systematic model comparison, evaluation of lumped model performance across UK regions, unseen validation, transferability to another country, and operationalization for early warning.

Accordingly, we revised the Aims and Objectives section (Lines 148–170) and the Methodology section introduction (Lines 170–187), adding the updated Figure 1.

Data statement and line 102

The latest (and likely last) version of the EDII is available with doi and should be cited as:

Blauhut, V., Stephan, R., and Stahl, K.: The European Drought Impact report Inventory (EDII V2.0), [data set], Uni Freiburg, Freiburg, https://doi.org/10.6094/UNIFR/230922, 2022

This is our preferred reference, because the website that is given by the authors no longer functions correctly.

>> We thank the reviewer for this suggestion. The citation of the EDII dataset has been updated as recommended, and the link in the Data Availability section has been corrected accordingly.

I don't see the reason to cite the new EDID database in the data statement as it was not used. Please consider that while EDID ingested a major part, but not all content of the former EDII, along with other databases, it uses different categories and different attributes than EDII. Therefore, naming it here and in lines 102 as one and the same is misleading as using it for the same purpose might provide different time series of NI etc.

For your information: A paper on the new db is in progress and about to be submitted to NHESS. Futhermore, guidelines for interested contributors on how to transfer EDII-categories into the new EDID-systems are already available:

Szillat, K., Hlavsová, M., Rossi, L., Blauhut, V., Stahl, K.: Transformation of text-based drought impact data from EDII (European Drought Impact report Inventory) to EDID (European Drought Impact Database): Guidelines. Freiburg HydroNotes no. 8. https://doi.org/10.6094/UNIFR/271380, 2025

>> We thank the reviewer for this comment. After careful consideration, we have removed the EDID reference and citation from the text, as it was not directly used in our study.

Minor comments

"Short term" = 0-3 months? For weather/floods anything more than a few days would be considered long term and not short term. Discussion and use of terminology might be improved.

>> We thank the reviewer for this comment. To avoid ambiguity, we have replaced the term "short term" in the manuscript and now refer explicitly to forecasts of up to three months for drought impacts throughout the main text.

line 13 'Here, "we" used ...?

>> The sentence is corrected, by adding "we".

"Here, we used data from the European Drought Impact Inventory ..."

Line 344: what is meant by 'regional information'? The impact occurrence? Why not use 'predictor' and 'predictand' or 'response' or so or better even, a variable name.

>> The term "regional information" has been replaced with the specific variables "AAR and PopR" in Line 354 to improve clarity.

"... while AAR and PopR were kept constant for all grid cells within each NUTS1 region."

line 674. EDID is not global. Replace with something global or say 'Europe'.

>> EDID is removed from the text.

Figures

Figure 5 what are dashed and solid lines? A legend with symbols and line types is strongly preferable over the difficult to read/miss caption text.

>> We thank the reviewer for this comment. The dashed lines indicate training results, while the solid lines show validation AUC results. We agree that the previous caption was difficult to read, so in the revised manuscript we have added a clearer legend explanation.

Line 490-493: *"Figure 5 Impact prediction and forecast accuracy (AUC) of the RF model for the UK in binary (squares) and categorical severity (circles) classifications. The x-axis shows lead times* *from prediction ("Pred.") to 24-month forecasts, and the y-axis shows AUC. Colours indicate model configurations with different predictor sets (top-right legend), with the black line using all predictors. On the legend at top left, dashed lines represent training performance and solid lines represent validation performance."*

Also, in Figure 2 it is confusing that some components have a legend and others don't. It should at least be consistent.

>> We thank the reviewer for pointing this out. In the revised manuscript, we have ensured consistency across all components of Figure 2, so that every element is either clearly labelled or accompanied by a legend, making the figure easier to interpret.

Line 409-413: *"Figure 2 Observed and predicted number of impacts (NI), severity categories, and likelihood of occurrences for the NUTS1 London region. In all panels, the positive y-axis shows observed values and the negative y-axis shows predicted values. Panel (a) shows NI values (square root transformed) from the Random Forest model, with the blue line marking the threshold below which NI is set to zero. Panel (b) shows severity categories (tercile method), and panel (c) shows likelihood of occurrence. Grey shading indicates the independent validation period (2007–2012)."*

The caption in Fig. 11 needs to name all panels (or a to d…etc..) but just singling out some is inconsistent.

>> We thank the reviewer for this comment. The caption of Figure 7 (which corresponds to the intended figure, as there is no Figure 11) has been revised to clearly name all panels (a–p) and describe their contents.

Line 546-550: *"Figure 7 Observed NUTS1-level and predicted/forecasted gridded model results over the UK for July 1984 and April 2012, with forecast lead times (LT) of 0 and 3 months. Panels (a–d) show July 1984 at LT0 (observed, predicted, severity, likelihood), panels (e–h) show July 1984 at LT3, panels (i–l) show April 2012 at LT0, and panels (m–p) show April 2012 at LT3. Within each group, panels are ordered as: observed occurrences ($\sqrt{NI}$), predicted occurrences ($\sqrt{NI}$), severity categories, and likelihood of occurrences."*