Response to R #2

We thank the reviewer for their constructive comments and the time that they invested in reviewing the manuscript. Our responses are included in blue text below and proposed additions are included in red.

Review comments for “What is a drought-to-flood transition? Pitfalls and recommendations for defining consecutive hydrological extreme events” by Anderson et al.

The present study provides a timely and well-motivated investigation into the definition and identification of drought-to-flood transition event, which is an increasingly important topic in hydrology. By comparing multiple threshold-based methods across eight case study catchments, the authors demonstrate how methodological choices can significantly influence event detection and interpretation. These findings carry potential implications for understanding compound and consecutive hydrological hazards. However, several methods appear to introduce potential biases. In addition, the study lacks a strong physical grounding, and certain sections would benefit from clearer justification and interpretation to enhance scientific rigor and applicability. Please see my detailed comments below:

We thank the reviewer for their constructive comments and agree with many of their remarks. The point that the study lacks a strong physical grounding is, to some extent, the one that we are trying to make with reference to recent work on the subject.

We use methods which are popular in this growing field, as well as in drought and flood research more generally. One of the primary take-aways from our paper is that using these methods without carefully accessing their physical basis blurs the meaning of drought to flood transitions research because many different event detection methods are applied across studies, and, thus, are not easily interpretable in the context of impacts or change in physical response. It is for this reason that our qualitative approach introduces the novel idea of using “meaningful impacts” as benchmarks (see response to similar comments by R#1) . Further, we use eight real-world case study events to assess the validity of existing approaches.

We address the comments specifically in the blue text below.

1. Abstract: The current abstract is primarily qualitative in nature and may benefit from the inclusion of key quantitative findings to strengthen its scientific clarity.

   We will include quantitative examples of event detection numbers between regimes and the number of detected case study events as below:

   “The number and timing of transitions differs substantially between threshold level approaches in highly seasonal regimes as opposed to those with a weaker seasonality.

   “We show that the probability of a transition occurring within a set time window could vary substantially between different methodologies and catchments.

   “For the eight case study events taken from media, governmental and scientific reports, only three of the transitions were successfully detected.”

2. L106-108: The reliance on streamflow thresholds without incorporating standardized indices may limit comparability with other studies and hinder the evaluation of meteorological or soil moisture-driven processes. The authors should more clearly justify this choice and discuss its implications for generalizability.

   We thank the reviewer for this comment. We originally performed the same analysis with the standardized streamflow index (SSI) defined using both fixed and rolling windows.

These approaches did result in different numbers, timing, and characteristics of transition events because of the effect of the aggregation.

Empirical percentiles can also be viewed as indices describing the historical non-exceedance frequency within a certain period as compared to the non-exceedance probability given by the SSI. As such, they are both indices that define deviations relative to what is the normal flow regime. Thus, they can be compared across different flow regimes (as well as other hydrometeorological time series). Using what is commonly referred to as standardised indices, such as the SSI, adds uncertainty as to the method used in its calculation, including choice of distribution and fitting method (parametric method) as well as choice of non-parametric methods (if preferred). SSI values (std. of the normalised distribution) have their counterparts in percentile values. The threshold level method, originally designed to be applied on raw streamflow data, has been applied to SSI time series using a given SSI value as threshold (e.g., −0.84 corresponding to the commonly used 20 percentile threshold).

In this study, focus is on hydrological extremes (i.e. drought to flood transitions) as these are extremes causing impacts on the ground. A meteorological extreme, e.g. SPI, does not necessarily translate to an impact on the ground (Brunner et al., 2025), this would depend on catchment characteristics (response time in particular) and wetness status. The study is motivated by the fact that floods and droughts are reported to occur more frequently and are often more extreme. In particular, extreme rainfall on dry grounds, may lead to extreme flooding, and one key question is whether reported extreme flood events can be linked to dry preconditions (i.e., drought to flood transitions).

By benchmarking our case study examples against reported impacts, we introduce a novel approach based on independent data sources. Despite these being media reported impacts that come with some uncertainty to the timing and areal coverage, they provide valuable insight into the relevance of the indices evaluated, notable at the local scale. Society and users are primarily interested in indices that can translate into impacts, which is also an approach gaining increasing application around the world (referred to as Impact-based forecasting or prediction) and among others recommended by WMO. We will make this aspect and motivation clearer in the Introduction.

We will add 2 additional sentences to clarify our reasoning for excluding these in the text as follows:

“The choice to look at events only, rather than periods of anomalous flow, as with standardized indices, was made because we are interested in hydrological extremes and standardized meteorological extremes may not translate to hydrological response (Brunner et al., 2025). Further, because the use of standardized indices such as the standardized streamflow index (SSI) introduces uncertainty to the method used in its calculation, including choice of distribution and fitting method (parametric method) as well as choice of non-parametric methods (if preferred).”

3. L140-143: The approach of defining drought and flood periods based on “reported” start and end dates from media or scientific sources may introduce a degree of subjectivity, particularly as media reports can vary widely in temporal resolution and geographic specificity. Did the authors use any verification measures to address potential reporting bias or spatial/temporal mismatches? It is also recommended that these reported dates be supplemented with objective hydroclimatic metrics to enhance consistency.

A great deal of time and effort went into validating that the catchments used in the analysis were appropriate to the reported events, Own experience from previous work collecting drought related impacts (e.g. Stahl et al., 2016), generally finds that media reports are reliable and valuable sources of information as to when and where impacts

occur. As for drought, the full spatial extent may not always be assessed given the local focus on many media outlets (e.g. local newspaper). In the case of floods, the media may not report the exact day of the peak discharge, but the actual impacts of the high discharge on the ground. The details of the information provided is considered valuable in particular at the local scale as this level of detail is seldom reported by state or regional authorities. Many candidate case study events were considered and the final eight were ultimately selected because they had a combination of reported impacts, and sufficiently long records of corresponding streamflow data.

As described in the text, a 10-day buffer period was used to allow for variation in the flood timing to accommodate for temporal mismatch. The occurrence of the reported flood events was compared with the peak discharge in the time series of every case study catchment, except for the Emme in Switzerland. This is demonstrated in Figure 3. The flood on the Emme was not detected in the daily streamflow timeseries as it was a rapid-developing flood. It was therefore validated using the hourly time series as reported in the text and in Figure 4. Validating that catchments fell within an area of hydrological drought was more challenging because reporting times, start and end dates, and regions can vary widely. Further, drought types are often conflated in media reports, and many different streamflow thresholds are used in scientific literature.

However, all case study drought periods were at least partially detected by a minimum of one method, with the exception of the drought in California. In this case, drought clearly documented throughout the region (DeFlorio et al., 2023), including low river and lake levels and water shortages as reported by the local government. While low water levels are reflected in the 2020-2022 period for the Ventura River, they are never zero. One news source speculated that this was possibly due to human influence and the importance of the river for city water provision (Faraday, 2021), noting that pumping from the river was halted in September 2021 to prevent it from running dry. This, and prolonged, intense, multi-year drought earlier in the time series are likely the reason that the drought period was not captured here.

The following text will be added to the qualitative assessment section to highlight this important point:

“Further, whether or not a drought to flood transition is apparent in the streamflow data, can also be influenced by human behavior. For example, the Californian case study (Ventura River) provides a substantial proportion of the water supply to the local municipality (Walter et al., 2015). The river is managed to sustain flows during prolonged dry spells via the Casitas Reservoir since cycling between drought and flood periods is the norm in the catchment (Walter et al., 2015). One news source noted that the local government halted water extractions in 2021, during the case study drought to flood period, thus preventing the river from running dry (Feraday, 2021). Management strategies like these mean that, without place specific knowledge, drought, and therefore also drought to flood transitions, may be overlooked in this, and in many other, managed catchments globally.”

In all cases, more than one reference was found for the event, and the earliest listed start date and latest end date (for drought) were used. The occurrence of a flood is best validated by the fact that floods were detected in every time series. However, in initial scoping phases, where available, satellite derived flood polygons were overlaid with the catchment areas to validate that the catchment fell within the flooded zones for specific events (e.g. in Italy https://mapping.emergency.copernicus.eu/activations/EMSN154/#activation-deliverables and Norway https://data.jrc.ec.europa.eu/dataset/1bd644a5-41cc-42c3-aff8-30dc7a8b1bc9). Wherever possible, events were validated using government reported data with reference to specific catchments (e.g. in Switzerland

https://www.bafu.admin.ch/dam/bafu/de/dokumente/hydrologie/uz-umwelt-zustand/hydrologisches-jahrbuch-2022.pdf.download.pdf/de_BAFU_UZ_2215_Hydrologisches_Jahrbuch_2022_bf.pdf, California https://www.cnrfc.noaa.gov/storm_summaries/dec2022Jan2023storms.php, Australia http://www.bom.gov.au/qld/flood/brochures/daintree/daintree.shtml, and Texas for drought https://www.lcra.org/news/news-releases/no-highland-lakes-water-available-for-lcra-agricultural-customers-this-year/) or regions (Norway for drought https://www.nlr.no/files/documents/NLR-SA/2024/Ostlandet/Torkesommeren-2018.pdf, and Texas for drought https://www.lcra.org/news/news-releases/no-highland-lakes-water-available-for-lcra-agricultural-customers-this-year/).

To address this concern, we will include a more detailed description of the validation approach in the manuscript including the following text:

Media reports may be misleading in terms of event location of timing. We aimed to validate event occurrence and the correspondence of spatial area by allowing for buffer periods around flood events, confirming the presence of a flood on the expected dates in the time series of all cases, relying on more than one report for all events, seeking scientific and governmental reports which listed specific catchment IDs rather than news media. Previous research has also suggested that media reports can be reliable and valuable sources of information as to when and where impacts occur \citep{stahl_impacts_2016}. References and descriptions of the events are listed in Table \ref{tab:drought_flood}. It is worth noting that validation in the case of drought was more difficult, given the generally wider spatial areas, and varied definitions used in reporting.

4. L161-162: This limited data coverage (17–20 years) may compromise statistical robustness of this study. Did the authors perform any sensitivity checks to assess the influence of record length on event detection and methodology performance?

The data coverage is only limited in 2 catchments.

As described in the text, six out of eight case study catchments had greater than 40 years of 95% complete daily streamflow data (96 years in the Californian catchment). We acknowledge that shorter time series effect the threshold-based event detection as fewer years may not capture the full breadth in hydrological variability as compared to longer time series. However, any potential limitations introduced by short data records would apply consistently across all methods (fixed and variable thresholds, for both drought and flood detection). Since the aim of this study is not to precisely quantify drought or flood frequencies and their transitions in each case studies, but rather to compare how different approaches define and detect drought-to-flood transitions, the relative comparison remains valid. In this context, the limited time series length may affect absolute estimates but does not affect the comparative outcomes across methods, which are central to the study's objectives. We will add the following sentence: "The use of fixed or variable threshold analysis on short time series may not capture the full range of variability in the streamflow as compared to longer time series. However, this would apply consistently across all threshold level approaches and methodological combinations (fixed and variable thresholds, for both drought and flood detection)."

5. L169-174: Here, the rationale for choosing the specific percentile window (±15 days) and the smoothing approach (31-day rolling mean) may need further justification.

We will add the following justification:

"This 31-day window was selected because it offers sufficient smoothing to capture seasonal behaviour without over-smoothing, representing the monthly time scale (Van Loon et al., 2012, Tallaksen and Van Lanen, 2023)"

And for the 30 days prior used in smoothing streamflow for drought detection:

"The previous 30-days are used, rather than a centered period, so that rapidly occurring flood events do not bias the drought end date."

6. L194-196: Calibrating threshold percentiles to match target event frequencies may obscure the physical hydrological meaning of the thresholds. These calibrated percentiles may not align with widely accepted definitions of drought and flood conditions, potentially limiting comparability with other studies. The reviewer suggests that the authors further discuss its potential biases and implications for physical interpretation.

Statistical threshold levels are rarely directly tied to clear physical processes. Although we acknowledge that the calibration approach used here is unusual, the threshold levels chosen for drought (ranging between 0.12 and 0.15) fall within the (stricter end of a) range of commonly used hydrological drought thresholds. Further, this approach allows us to fairly compare the methodologies by selecting the same number of events, as described in the text. In this way, the bias of the threshold level is reduced, and the difference in the threshold type can be considered more directly. The 98th percentile threshold used for flood is also common, and approximates the 2-year recurrence interval, which is often substituted (although somewhat erroneously, Liu et al., 2025) for bankfull stage.

We do not believe that the selection of a 0.12 percentile threshold is any less physically interpretable than the selection of a 0.1 or 0.2 percentile threshold would have been. Barring the definition of threshold levels derived from local and/or impact specific knowledge (e.g. the water level at which energy production is negatively affected in an individual river, or the level at which a particular fish species is known to be threatened), we feel that this approach is already well justified in the text.

We will edit the current description to read: "Defining the threshold based on the number of events (selecting the same number of events) was done so that the different threshold level approaches could be compared more fairly. We repeated the event detection approach using a range of threshold levels until the desired number of events was selected regardless of methodology. Alternatively, the same percentile value could have been used for all methods. Threshold level methods may result in the selection of a proportion of streamflow which is not equivalent to the stated percentile (Brunner and Voigt, 2024), in part because of the imposed minimum duration for drought periods (30 days). Although this may not have given the same number of events, it would have provided the same number of days below the threshold."

7. L207-209: The fixed 90-day interval for defining transitions oversimplifies the wide range of hydrological responses across diverse catchments and climatic regimes. The authors should justify the applicability of this threshold across all case studies and conduct sensitivity analyses to assess how the chosen interval affects event detection.

We completely agree! We use the 90- and 14-day thresholds because these, and similar windows, were used in previous research. Figure 8.b. is intended to address this important point. We will include additional discussion of this point following proposed changes to the figure, as described in the next response.

8. Line 230: The choice of the 5th percentile as a transition threshold lacks clear justification. Besides, while bootstrapping accounts for sparse or uneven data, it may also smooth over physically meaningful variability between catchments. The reviewer recommends discussing

the sensitivity of results to this resampling approach and analyzing how catchment-specific hydrological characteristics influence the derived thresholds.

Since the submission of the manuscript, we have reconsidered this approach. In the resubmission, we instead propose to include a simpler and more direct assessment of the time windows between events. We have fitted a theoretical distribution (GEV) to the distributions of time periods between drought and flood. This allows us to avoid the bootstrapping procedure entirely. From this, we select the percentiles of time periods in each catchment corresponding to the 14- and 90-day time windows. We will include the updated figure, the following description of the methodology, and the subsequent additional discussion:

"Second, we consider how defining case specific time intervals between drought and flood events may affect the analysis. (1) We begin by calculating the time interval between each drought event and the first subsequent flood event, resulting in a series of time interval values for each catchment. (2) Next, we fit a GEV distribution to the time interval data series using the R package *extRemes* and use the GEV parameters to compute the cumulative distribution function (CDF) for a theoretical period of 0-730 days (2 years). Several candidate distributions were tested, and it was shown that the GEV distribution was a good fit (based on the Kolmogorov-Smirnov and Anderson-Darlings tests, for all case study catchments). This step is necessary because the events are sometimes unevenly distributed or too sparse to reliably estimate time interval probabilities from the empirical distribution directly. (3) From the CDF for each catchment, we extract the probability of 14- and 90-day transition periods."

The presentation in the discussion will be changed as follows:

"This approach also does not consider how time intervals between drought and flood are likely to differ under different hydrological conditions within the catchments. In some locations, the shift between dry and wet conditions may be rapid developing or span of several weeks and in either case be part of a normal seasonal pattern to which local populations and systems are adapted. As an alternative to the top-down selection of time intervals, we suggest that these should be defined based either on impact-specific needs, when possible (e.g. the minimum time frame for changing a reservoir management protocol), or be defined relative to what is normal conditions if this is not feasible.

Here, we tested an alternative approach in which the time intervals between all drought events and the first subsequent flood period were defined probabilistically (Figure 8.b.). The results indicate that, depending on the methodological approach and the flow regime, the probability of a transition can vary widely. For example, in the Norwegian case, the probability of a 14-day time interval between the end of drought and start of flood ranges from 0.02% when fixed thresholds are used (f_f), to 6.5% when variable thresholds are used (v_s). On the other end, the probability of a transition within 14 days in the Swiss case study catchment reaches as high as 14.5% (s_f) and has greater than a 10% chance of occurrence within this time window, regardless of the approach. These results show that the few transitions identified within the selected time windows may differ substantially among the case study catchments, which point to the need for more research on how to best define robust and meaningful transition schemes for hydrological extremes."

9. L235-236: The reviewer suggests the authors clarify the criteria used to distinguish between transitions "not of interest" and those "potentially representing negative impacts." This distinction appears subjective and would benefit from being more explicitly linked to hydrological or societal impact indicators.

We will rephrase this statement as follows:

“The motivation for studying drought to flood transitions ranges from changes in physical processes, like soil water repellency resulting from dry periods, to water management or infrastructural concerns, like reservoir operations (Hammond et al., 2025). In this manuscript, we are interested in understanding and contextualizing which methodological approach is able to detect transitions that are likely to have certain impacts or result from process changes that can be meaningfully interpreted in the desired context.”

Following visual inspection, we provide examples and demonstrate potential pitfalls in the typical event detection approaches

10. L391-394: While the authors stress methodological influence, it could more clearly relate detection differences back to physical processes such as antecedent soil moisture and rainfall intensity to aid in bridging the gap between statistical detection and hydrological realism.

We thank the reviewer for this comment. In our view, it will be difficult to clearly map detection differences at the event scale onto physical processes or event characteristics. These differences are more likely related to the overall streamflow regime. This comment is partially addressed by our previous response highlighting the difference between threshold level methods and how they relate to seasonality in particular. We will additionally add some further information regarding how certain methodologies are favor different drought and flood types, e.g.

“Thus, these different threshold types may result in the detection of different types of flood and drought events. Fixed thresholds will detect those which are more in line with the seasonal norms, e.g. flood events which occur during the high flow season and droughts which occur during the low flow season. These events are more likely to be driven by larger scale processes, for example, drought periods driven by cold winter weather or floods driven by snow melt in the spring, although very severe events would be detected outside of seasonal patterns. In contrast, variable thresholds are sensitive to anomalies within a season. This means they can capture atypical events, which may not necessarily be large in terms of absolute magnitude, for example, droughts during typically wet periods which may be indicative of short-term rainfall deficits or a delay in the snow melt season, for example. Floods events during drier periods potentially driven by intense, short-duration rainfall which does not result in a large absolute magnitude of flow, could also be detected. In other words, variable thresholds may detect events which are not extreme in an absolute sense, but which are anomalous in a specific time of year. In regimes which do not have a strong seasonal cycle, the threshold choices should not result in the detection of substantially different events.”