

RESEARCH ARTICLE OPEN ACCESS

How Good Is Your Forecast? The Importance of User-Centred, Decision-Relevant Evaluation of Forecasts and Climate Services

Rebecca Emerton¹  | Schalk Jan van Andel²  | Katherine Egan¹ | Calum Baugh¹  | Deborah Dotta Correa^{2,3}  | Emie Klein Holkenborg⁴ | Daniele Castellana⁴ | Marc van den Homberg^{4,5}  | Yiheng Du⁶  | Ilias Pechlivanidis⁶  | Sumiran Rastogi²  | Micha Werner² 

¹European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK | ²IHE Delft Institute for Water Education, Delft, the Netherlands | ³Wageningen University, Wageningen, the Netherlands | ⁴The Netherlands Red Cross, Den Haag, the Netherlands | ⁵University of Twente, Enschede, the Netherlands | ⁶Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden

Correspondence: Rebecca Emerton (rebecca.emerton@ecmwf.int)

Received: 9 December 2025 | **Revised:** 15 July 2026 | **Accepted:** 3 August 2026

ABSTRACT

Forecasts and climate services, spanning weather and climate timescales, play a vital role in supporting short- and long-term decision-making across a wide range of climate-sensitive sectors. Understanding their skill is crucial to ensuring that the information provided is relevant, trustworthy and usable in real-world contexts. Traditional evaluation and verification methods typically focus on statistical accuracy, which remains essential for model development and improvement. However, these metrics alone may not capture the aspects of forecast performance that matter most to users and often fail to reflect how well forecasts could support practical decision-making. This paper highlights the importance of user-driven, decision-relevant evaluation: an approach that centres on the needs, experiences and decision-making contexts of forecast and climate service users. We present a flexible framework for user-driven forecast evaluation. Drawing on case studies from diverse regions and hazard types—including cold waves in Lesotho, precipitation variability in Spain and drought in Lesotho, the Netherlands, Italy and Georgia—we demonstrate how user engagement can tailor evaluation to decision-making contexts and local knowledge. The framework combines event-based contingency table approaches, using user-defined thresholds that are linked to the decision they make, to derive transparent metrics such as hits, misses and false alarms, alongside scores such as probability of detection, false alarm ratio and threat score. Traditional verification metrics (such as mean absolute error or the continuous ranked probability score) are applied as complementary diagnostics, while co-evaluation of forecast performance for historical events is used to further support dialogue, accessibility and trust. The case studies show the value of combining long-term performance metrics with context-rich narratives and emphasise the importance of data visualisation and communication in shaping forecast usability. We conclude with recommendations for integrating user-centred evaluation into operational workflows to ensure climate services are both scientifically robust and decision-relevant.

1 | Introduction

Climate services play a vital role in supporting decision-making across a wide range of climate-sensitive sectors, from disaster risk reduction to water resources and agricultural management

(Vaughan and Dessai 2014; Lemos and Morehouse 2005; White et al. 2022).

A climate service is broadly defined as the transformation of climate-related data—together with relevant knowledge—into

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 European Centre for Medium-Range Weather Forecasts (ECMWF), Netherlands Red Cross and The Author(s). *Meteorological Applications* published by John Wiley & Sons Ltd on behalf of Royal Meteorological Society.

customised products and actionable knowledge by providing tailored information, tools and platforms that guide users towards climate-resilient choices (Crochemore et al. 2024; Cantone et al. 2023). These products could be forecasts, warnings, projections, trends, risk assessments, reports or other resources that provide information in relation to climate that may be of use for society (European Commission: Directorate-General for Research and Innovation et al. 2015; Wens et al. 2025). In this paper, we use the term ‘climate services’ to encompass information products spanning both weather and climate timescales, ranging from short- and medium-range forecasts to seasonal forecasts.

Co-creation is increasingly recognised as a preferred practice in climate service development (McClure et al. 2024). It is an inclusive, interdisciplinary, interactive and iterative approach to developing climate services that aims to overcome divisions between scientists, developers and decision-makers by fostering shared ownership and trust (Lemos and Morehouse 2005; Nyamakura et al. 2025). We use the term co-creation to describe the collaborative process encompassing the co-design, co-production, co-implementation, co-evaluation and dissemination of user-centred services. Such a framework promotes climate services that are tailored to contextual realities and capable of meeting evolving needs (Bamzai-Dodson and McPherson 2022).

Understanding the skill and societal value of forecasts and climate services is crucial to ensuring that the information provided is not only scientifically sound but also relevant, trustworthy and usable in real-world contexts (Pechlivianidis et al. 2025). Co-evaluation is one step in the process of co-creation of a climate service, which ensures that this process is informed by users’ local knowledge and perspectives. The usability of a climate service can be evaluated from the perspective of salience (decision relevance, whether the information is actionable and accessible), legitimacy (whether the user is fairly represented with an understanding of how and why the information was produced) and credibility (underpinning both the salience and legitimacy, whether the information is authoritative, believable and scientifically sound within the limits of scientific predictability and uncertainty), including aspects such as the delivery mechanism, visualisation, interpretation or decision support. These should be considered holistically in the co-creation of a climate service. In this paper, the focus is on evaluating the skill of forecasts in a way that is user-centred and decision-relevant and in collaboration with decision-makers.

The I-CISK project (Innovating Climate Services through Integrating Scientific and Local Knowledge; www.icisk.eu) has worked to co-create (Wens et al. 2025) human-centred climate services in collaboration with a range of stakeholders across a variety of different sectors in Europe and Africa (Masih et al. 2022).

In the following sections, we demonstrate a framework for user-driven forecast evaluation and discuss experiences and insights from case studies conducted in real-world environments that have followed this framework within the I-CISK project before presenting key conclusions and recommendations.

1.1 | How Does User-Driven Evaluation Differ From ‘Traditional’ Forecast Evaluation?

Traditional forecast evaluation, or verification, refers to the process of assessing the performance of forecasts by comparing them to observed outcomes using statistical measures. This can include a wide range of different metrics, encompassing both deterministic scores (such as the mean error, correlation or bias) and probabilistic measures (such as ROC analysis based on contingency tables or distribution-based scores such as the Continuous Ranked Probability Score [CRPS]), each assessing one or several forecast quality attributes (Murphy 1993; WMO 2025). Skill measures the relative accuracy of a forecast compared to a benchmark, quantifying how much more (or less) informative a forecast is than the benchmark, such as climatology, persistence or another forecasting system (Pappenberger et al. 2015). These objective methods are essential for assessing the technical performance of forecasting systems, understanding their accuracy and driving model development. It is recognised, however, that such measures of forecast skill are not sufficient to determine usefulness for decision-making, as the value of a forecast depends on how it intersects with user decisions and thresholds rather than accuracy alone (Richardson 2000).

User-centred, decision-relevant evaluation assesses performance relative to user-defined decision thresholds (e.g., rainfall above a level that triggers flooding or temperature extremes affecting crops at a certain time of year) (Pagano et al. 2024). This type of evaluation is essential because a forecast that is statistically accurate according to a standard set of traditional evaluation metrics may nevertheless exhibit poor performance under the specific criteria relevant to a given user. In addition, it can be unclear how to use traditional evaluation metrics in an operational, decision-making environment. Incorporating user perspectives helps ensure forecasts are both robust and actionable. User-driven evaluation can also provide valuable insights for model development, as it may highlight behaviour in the extremes of the distribution or reveal systematic tendencies that may be obscured by aggregate verification metrics. In recent years, the forecast evaluation and verification community has increased emphasis on developing and implementing user-driven forecast evaluation (e.g., Ebert et al. 2018; Crochemore et al. 2024), yet more remains to be done to ensure that user needs and decision contexts are consistently and meaningfully integrated into operational evaluation practices.

2 | Evaluation Framework and Methodologies

2.1 | User-Driven Evaluation Framework

As part of its aim to advance more human-centred climate services, the I-CISK project developed a flexible, participatory framework for user-driven evaluation. This approach was designed to assess the relevance, usability and decision-support value of climate services from the perspective of those who use them. While not intended as a universal or definitive model, the framework offers one example of how user-centred evaluation can be structured and applied in practice. It was co-developed within the project and adapted to the specific contexts of I-CISK’s ‘Living Labs’.

Framework for user-driven evaluation

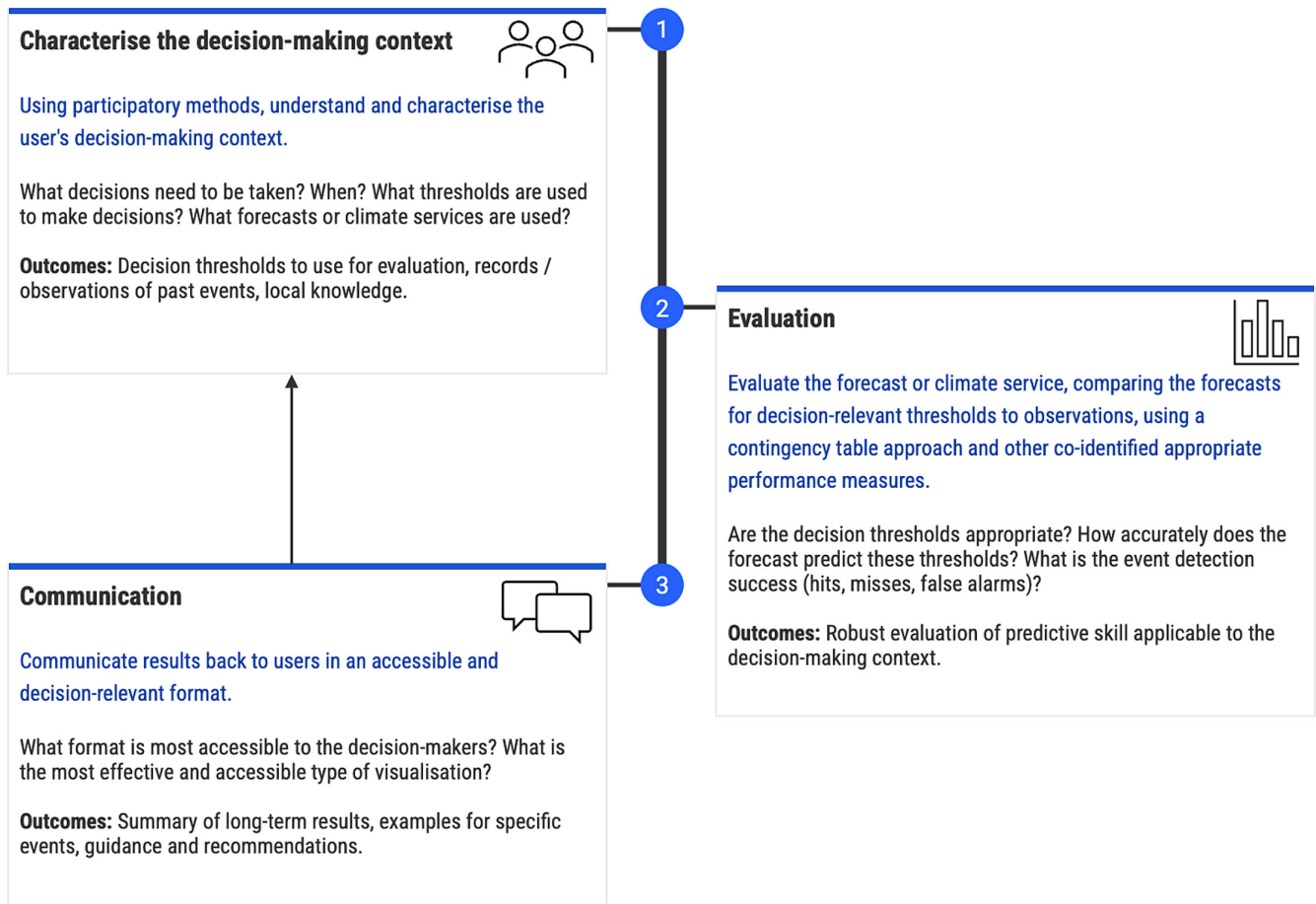


FIGURE 1 | Overview of the user-driven and decision-relevant evaluation framework used in the I-CISK project.

The framework consists of three core stages (Figure 1). First, the user's decision-making context is characterised through participatory methods. This step identifies what decisions are made, when those decisions occur, what thresholds are used to trigger action and whether historical records of relevant events or observational data are available. These thresholds serve as the basis for evaluating whether forecasts would have successfully informed action in past situations of interest. It can be challenging at first to identify these thresholds in some cases, and this process itself may need to be iterative and refined over time.

Second, the climate service is evaluated over a historical period by applying the user-centred and decision-relevant thresholds to forecasts and comparing the results to observations or reanalysis data. This is done using a contingency table approach, which summarises the number of hits, misses and false alarms. From the contingency tables, metrics such as probability of detection (POD), false alarm ratio (FAR) and threat score are computed. Evidence from the I-CISK project, spanning multiple Living Labs, found that contingency table categories, metrics and visualisations are both intuitive to users and directly applicable to decision-making (Baugh et al. 2025).

Third, results are communicated back to users in an accessible and decision-relevant format. Visualisations such as heatmaps,

bar charts and event comparisons are used to illustrate forecast performance. These products aim to support user interpretation by highlighting which lead times and forecast probabilities provide the most reliable guidance, under what conditions forecasts may be more reliable or more uncertain and how well the forecasts align with real-world impacts and user-defined thresholds.

Taking an iterative approach to each stage of the framework allows feedback from users to shape the methodologies, visualisation and interpretation, yielding a structured but flexible approach to climate service evaluation.

2.2 | Participatory Methods

In the context of climate services, participatory methods are approaches that actively involve users in the co-creation of climate services, ensuring that their knowledge, experiences and decision contexts shape both the process and the outcomes. A 'decision timeline' is a participatory method used to establish users' decision-making processes, cues and how these relate to climate service information (Calvel et al. 2020; van den Homberg et al. 2024). A decision timeline identifies the activities a user undertakes during the year, the climatic risks associated with these activities, the climate service information and/or local

Decision timeline for farmers in Los Pedroches, Spain

		Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Precipitation	Dairy		100 mm is ideal for germination of winter fodder crops		200–300 mm is ideal considering water availability for livestock								
	Dehesa		150–200 mm is ideal					400 mm ideal for winter crop					
	Olive	200 mm is needed for fertilisation and growth			A minimum of 300 mm of rain annually is sufficient, 500–600 mm annual is ideal								
Temperature	Dairy											Upper limit of 35–40°C	
	Dehesa				60 or more consecutive days with temperatures exceeding 32°C causes animals to suffer								
	Olive							No frost					
								Just one night with temperatures of -7°C or below can kill trees					
								Not >30°C				No weeks with sustained temperatures >40°C	
Agricultural activities		Deciding type of crop to grow in spring	Winter crop sowing				Flowering from an average of 15–20°C	Harvesting the seedbed					
Forecasts needed		How much will it rain in spring?											How much rain will fall in October?
													How much will it rain in Autumn?

FIGURE 2 | Example of a decision timeline produced as part of the I-CISK project for farmers in Spain, indicating key constraints related to precipitation and temperature conditions, such as ideal amounts of rain or challenging temperature ranges, alongside key agricultural activities and the timing required for forecast information.

knowledge used to anticipate such risks and can also consider the coping or adaptation strategies in place. The results are typically presented as a calendar showing the annual distribution and duration of each activity, risk and coping strategy. An example of a decision timeline, co-created in the Spanish Living Lab, is shown in Figure 2.

This method is valuable for identifying both existing and potential links between climate services and users' decision-making and local knowledge. Establishing these links is a critical step in evaluating the relevance and potential value of climate services in operational contexts and in narrowing the 'climate information usability gap' (Lemos et al. 2012).

Decision timelines are co-created during participatory workshops, with one timeline developed per user group (a group of individuals or organisations that carry out the same activities within a given area). For example, farmers cultivating the same crop in a region may be considered a single user group, as their decision-making processes are likely to be similar. It is important, however, that participants represent the diversity of experiences and perspectives within each group.

The process of creating a decision timeline typically follows four steps. (1) The user group first defines its sectoral activities and their timing throughout the year. For example, crop farmers in Spain identified some of their key activities: deciding what type of crop to grow in spring (September), sowing winter crops (October–November), flowering of the crops (February–March) and harvesting the seedbed (April–May) (Figure 2). (2) The group then defines any climatic and socio-economic risks associated with each activity. For example, for olive farmers, 200 mm of rainfall is needed in September–October for olive fertilisation and growth, while no precipitation in May–June is ideal and

temperatures below -7°C at any time of year pose a risk to the trees. Where possible, participants provide quantitative thresholds for these risks (e.g., daily, monthly or seasonal accumulated precipitation representing 'ideal rainfall'), which can later be used to evaluate the skill of relevant forecasts. Defining these thresholds can be challenging, particularly where users apply more qualitative descriptors than quantitatively defined thresholds. This may then require iteration to arrive at meaningful and decision-relevant values. (3) Next, the group identifies how these risks are currently monitored and anticipated; this may be through cues based on local knowledge (weather and environment or socioeconomic factors) and/or through forecasts and climate services. If the latter are not yet in use, the exercise and evaluation can help identify potentially useful forecasts or climate services for the decision-making context. (4) Finally, coping and adaptation strategies are documented where possible. For example, based on the forecast of how much rain is anticipated in spring, farmers may decide which crop to grow or to not grow any at all.

The completed decision timeline highlights where and how climate service information can support decision-making for each user group. It also identifies opportunities to evaluate forecast skill and usefulness in direct relation to user-defined activities, risks and thresholds.

2.3 | Evaluation Methodologies

2.3.1 | Contingency Table Approach

The framework for user-driven evaluation applies a contingency table approach, in which forecasts are assessed against user-defined thresholds and classified as *hits* (the event was

forecast and occurred in reality), *misses* (the event occurred but was not forecast), *false alarms* (the event was forecast but did not occur) or *correct negatives* (the event did not occur and was not forecast). This simple structure provides a direct and intuitive link between forecast performance and decision-making outcomes.

From these four categories, a range of evaluation metrics can be derived, as follows:

- *Probability of Detection (POD)* (also known as ‘hit rate’) measures the proportion of observed events that were successfully forecast. Calculated as $(hits/(hits + misses))$, a value of 1 (or 100%) means all events were detected. A high value is desirable, but this score does not penalise the forecast for false alarms, so a forecast that predicts an event every day could have a POD of 100%.
- *False Alarm Ratio (FAR)* indicates how often forecasted events failed to occur. Calculated as $(false\ alarms/(hits + false\ alarms))$, a value of 0 is perfect (no false alarms). A low value indicates a system that issues alerts only when necessary, but as missed events are not considered, a forecast that never predicts an event could have a FAR of 0.
- *Threat score* combines hits, misses and false alarms into a single measure of event detection, useful for judging overall reliability. Calculated as $(hits/(hits + misses + false\ alarms))$. Values closer to 1 (or 100%) are desirable, meaning the forecast captures events reliably while limiting false alarms or misses, depending on the tolerance for each in the decision context. However, values can still be low even for useful forecasts of rare events.

These scores capture different aspects of forecast performance and can be used together to highlight whether a system is more prone to missing critical events or issuing unnecessary warnings and to illustrate trade-offs between the two, which depend on the decision context and user priorities. Other scores also exist, but only those used in the case studies presented here are listed. In this study, contingency table-based metrics are used as transparent summaries of forecast performance at user-defined decision thresholds to support stakeholder dialogue and decision-making rather than as universal metrics for ranking or optimising probabilistic forecast systems.

While the contingency table approach can provide a clear and intuitive summary of decision-relevant forecast performance, it can be non-trivial to apply, with a number of decisions needing to be made as to how best to apply the approach depending on the decision context. In many cases, an event-based evaluation is most relevant, focusing on whether the onset or duration of a user-defined event was correctly forecast. This avoids inflating performance by counting many consecutive days within one prolonged event but raises challenges around defining when an event starts and ends. Daily or weekly applications may be more suitable where decisions are continuous in nature. Depending on the user needs, a ‘fuzzy’ evaluation can also be used, in which forecasts that predict an event within a small time window (e.g., ± 1 day of onset) are considered successful. A similar approach could also be applied to the event

threshold itself. In the contingency table approach, events are defined categorically (e.g., above/below a threshold), which can lead to sharp distinctions for values close to the threshold. In practice, however, impacts may vary more smoothly, and values close to the threshold may have similar implications. Methods that account for this, for example by introducing tolerance ranges or weighted scoring around thresholds, could provide a more nuanced assessment of forecast performance in such cases.

For probabilistic forecasts, ensemble probabilities can be converted to binary outcomes by applying probability thresholds, varying these systematically to explore different balances between missed events and false alarms. Receiver operating characteristic (ROC) curves, plotting the trade-off between detection rate and false alarm rate across these probability thresholds, can be used to visualise these results (note that the false alarm rate used in ROC analysis differs from the FAR introduced earlier; the false alarm rate is calculated as the proportion of non-events that were incorrectly forecast as events, whereas the FAR expresses the proportion of forecast events that did not occur). ROC-based information can be translated into decision-making support through the calculation of relative economic value, explicitly linking forecast performance to user cost-loss ratios (Mylne 2002; Richardson 2000). This approach can provide a transparent and intuitive measure of whether a forecast has value for a given user, expressed in terms of potential economic benefit.

However, the application of such frameworks requires users to articulate cost-loss relationships that are not always readily specified in the operational contexts considered here. In the Living Labs, users consistently found contingency tables and their associated counts of hits, misses and false alarms more immediately accessible and easier to relate to concrete actions. For this reason, ROC curves were treated as supporting diagnostics within the broader user-centred evaluation framework. Application of this framework to different contexts may identify alternative or additional diagnostics more suitable, depending on the decision-making context and engagement with users.

2.3.2 | Supporting Traditional Evaluation Metrics

Alongside the contingency table approach, a small set of traditional evaluation metrics was used in some cases to provide complementary information on forecast behaviour. These metrics are widely used in model evaluation and verification (Wilks 2011). These measures can help to provide additional context to explain patterns in the event-based scores and support discussions with decision-makers in some of the Living Labs around the relative strengths and limitations of different approaches.

The following metrics were considered in at least one case study presented in this paper:

- *Mean absolute error (MAE)* can be used to summarise the average magnitude of forecast errors compared to observations in the same units as the forecast variable. Values

Case study locations

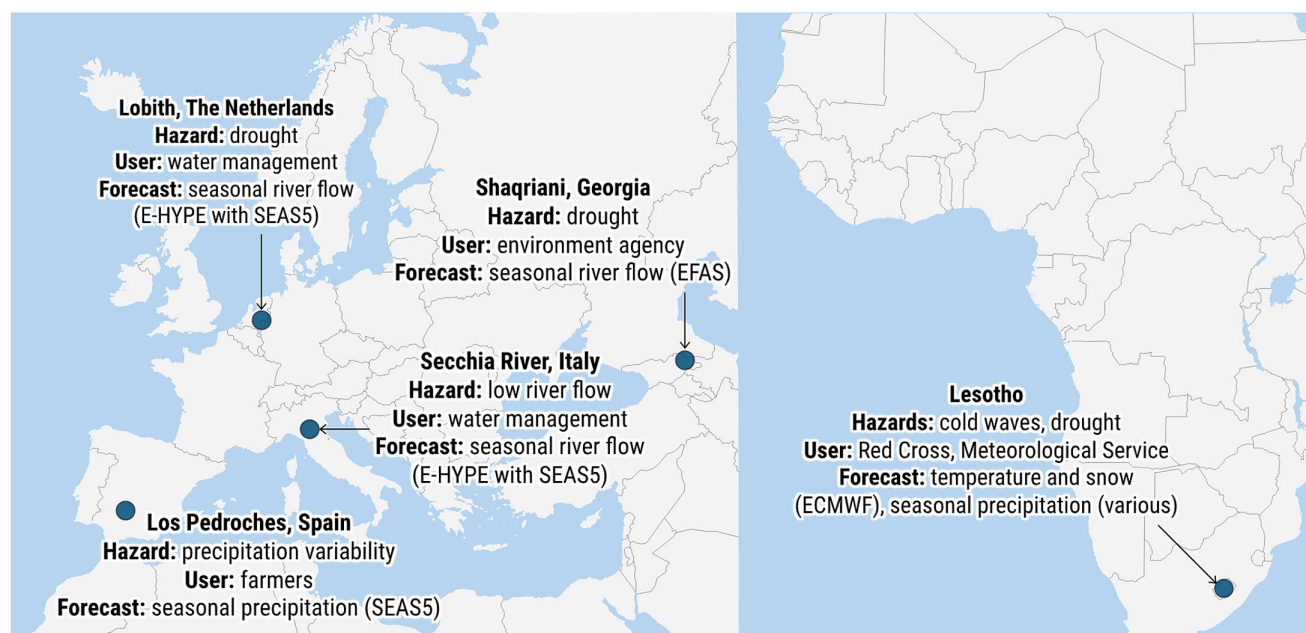


FIGURE 3 | Map of case study locations (I-CISK project ‘Living Labs’) for the user-driven, decision-relevant forecast evaluation, indicating the location, hazard of interest, key users involved in the co-evaluation and the type of forecast evaluated.

closer to 0 indicate smaller errors and therefore more accurate forecasts, but MAE does not distinguish between small frequent errors and occasional very large ones.

- *Mean bias* can be used to indicate systematic over- or under-prediction of the predicted variable in the same units as the variable. A value of 0 indicates no bias and smaller values are desirable.
- *Continuous Ranked Probability Score (CRPS)* (Matheson and Winkler 1976) can be applied to probabilistic forecasts to assess the accuracy of the full forecast distribution, relative to the observed outcome. Smaller values are better, with 0 indicating a perfect forecast. The CRPS is dominated by more frequent, “normal” conditions and may mask poor performance for extremes.
- The modified *Kling–Gupta Efficiency (KGE)* (Gupta et al. 2009) is commonly used in hydrology to evaluate continuous forecasts such as river discharge and can be used to combine correlation, variability and bias in continuous flow forecasts. Values closer to 1 (a perfect score) indicate forecasts that are well correlated with observations, with realistic variability and low bias, but the three components can compensate for each other (e.g., a good correlation may offset a strong bias).

While these traditional metrics remain essential for model evaluation and development, they do not on their own indicate whether a forecast is fit for purpose in operational decision-making. Their value here lies in complementing the contingency table results, offering a broader diagnostic picture while ensuring that evaluation remains relevant to thresholds and events of interest to decision-makers.

It is important to note that contingency table-based scores such as the threat score outlined in Section 2.3.1 are not proper scoring rules (i.e., scores that are optimised when forecast probabilities match observed frequencies) for probabilistic forecasts, as they rely on thresholding and therefore discard information about forecast probabilities. In this work, such metrics are not used to optimise or tune forecasts, but rather to summarise event-based performance once user-defined thresholds have been applied. For the development and comparison of probabilistic forecasting systems, proper scoring rules, such as CRPS or the Brier score, together with reliability diagnostics, remain essential complements.

3 | Case Studies of User-Driven Evaluation

In the following sections, we present a summary of results from the application of this framework in selected I-CISK Living Labs (Baugh et al. 2025): Lesotho, Spain, the Netherlands, Italy and Georgia. Figure 3 provides a map of the locations and key information such as the hazard of interest, key users involved in the co-evaluation and type of forecast evaluated.

3.1 | Anticipatory Action in Lesotho

Many humanitarian and governmental agencies implement anticipatory action (or early warning early action) to trigger early action in advance of hazardous weather events (Coughlan de Perez et al. 2015; Harrowsmith et al. 2020) to reduce the impact (Anticipation Hub 2025). To do so, early action protocols (EAPs) are co-created. These contain a trigger model, a set of pre-defined early actions and pre-agreed

Forecast timing accuracy for cold wave onset in Lesotho

Counts of forecasts, issued 1–14 days in advance, that predicted cold wave onset to occur earlier or later than observed

	1 day early/late	2 days early/late	3 days early/late
Predicted onset earlier than observed			
Lead time 1–3 days	10	1	
Lead time 4–6 days	10	1	1
Lead time 7–10 days	7	2	3
Lead time 11–14 days		5	
Predicted onset later than observed			
Lead time 1–3 days	1		
Lead time 4–6 days	1		
Lead time 7–10 days	3	2	
Lead time 11–14 days	2		

A cold wave is defined as ≥ 2 consecutive days with a maximum temperature $\leq 2^{\circ}\text{C}$ and snow accumulation ≥ 0.5 cm in the Lesotho Highlands. Based on ECMWF forecasts (2004–2022).

FIGURE 4 | Overview of the accuracy of predictions of the start date of a cold wave in ECMWF control (ex-HRES) re-forecasts, at lead times of 1–14 days, covering April–October 2004–2022.

funding for those actions. User-driven and decision-relevant forecast co-evaluation in the Lesotho Living Lab focused on analysing the performance of forecasts that are or could potentially be used within EAPs developed by the Lesotho Red Cross Society (LRCS) and Lesotho Meteorological Service (LMS) for cold waves and drought. An overview of each is provided separately below.

3.1.1 | Cold Waves

Freezing temperatures and snow characterise cold waves in Lesotho, posing a significant risk to communities and infrastructure. The EAP for cold waves (Anticipation Hub 2024) defines thresholds, based on forecasts, that trigger certain anticipatory actions. For example, distribution of clothing and cash transfers will be undertaken ‘when LMS issues a forecast of 2°C maximum day temperature or less in the Highlands for at least two consecutive days with the possibility of moderate snowfall in the Highlands’. The anticipated lead time for this type of action is in the range of 4 days.

The EAP is based on LMS forecast guidance, which itself is informed by a range of information, including forecasts from ECMWF. The evaluation aimed to support LMS and the EAP by assessing the ability of ECMWF medium-range forecasts (ECMWF 2025a) to predict cold wave events. Evaluation focused on the April–October cold season from 2004 to 2022, using re-forecasts (IFS cycle 48r1; ECMWF 2024) at 9 km resolution. Further detail on the user-driven evaluation methodology and the full set of results for both the deterministic (control) and ensemble forecasts can be found in Egan et al. (in preparation). A summary is provided here.

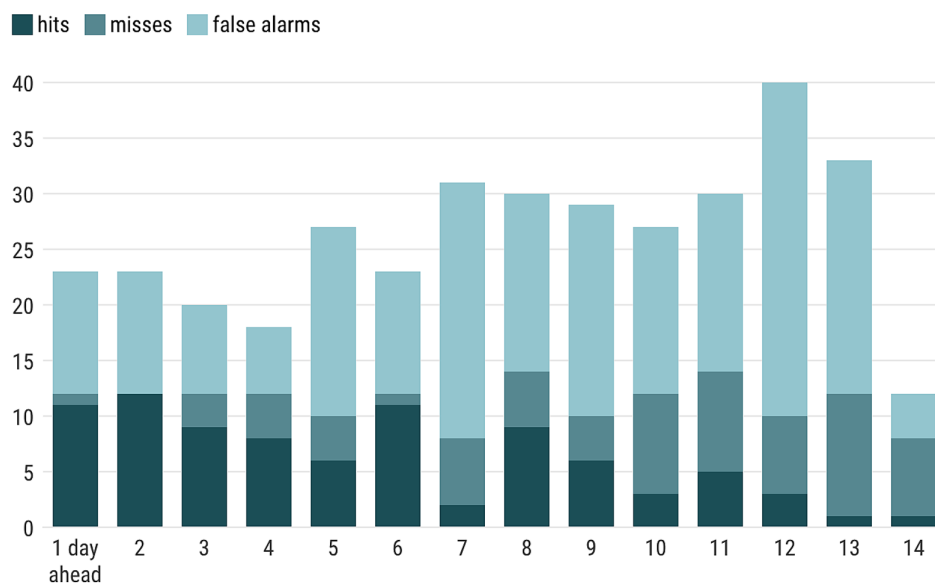
Sparse observational records, particularly for snow, posed a challenge for the evaluation. ERA5-Land reanalysis data

(Muñoz-Sabater et al. 2021) were therefore used as a proxy. Consistency of the reanalysis with observed events was ensured through the creation of an ‘impacts database’, which was used to tailor the EAP thresholds for use with gridded weather data. The database contained 21 cold wave events with documented impacts identified from institutional records, academic literature and media reports. These were compared, using a contingency table approach, to events identified in ERA5-Land by applying various EAP-like thresholds and selecting those that best captured the events in the impacts database. The tailored thresholds identified using this method closely resembled and further quantified the original EAP criteria: two or more consecutive days with a maximum temperature $\leq 2^{\circ}\text{C}$ and snow accumulation ≥ 0.5 cm in the highlands. A total of 40 events were identified in ERA5-Land using these criteria (the larger number of events compared to the impact database is expected due to the incomplete nature of the records used to construct the database). This approach successfully ensured that the evaluation was linked to real events by using records of past events, despite a lack of observed data.

An event-based, time-flexible application of the contingency table approach was used to investigate the ability of ECMWF forecasts to identify the start dates of these events identified in ERA5-Land. LRCS highlighted that for operational resource planning, it is critical to know the exact start date of a cold wave. Evaluating how early or late the forecasts tend to predict the start of an event provided useful information for an important aspect of decision-making, though forecast performance was suboptimal when required to predict exact start dates. Aggregated over lead times of 1–14 days, in total the deterministic forecasts produced 53 hits (correctly predicted onset dates), while 40 predictions were too early and 9 too late. Figure 4 indicates the number of predictions with a cold wave onset date up to 3 days earlier or later than the observed start of the event. Beyond 3 days, the forecast was considered to have missed the event, rather than a so-called ‘near miss’ for an event within ± 3 days of the predicted

Forecast evaluation for cold waves in Lesotho

Number of hits, misses and false alarms in ECMWF forecasts of cold waves from 2004 to 2022, at lead times of 1 to 14 days ahead



A cold wave is defined as two or more consecutive days with a maximum temperature $\leq 2^{\circ}\text{C}$ and snow accumulation ≥ 0.5 cm in the Lesotho Highlands. A hit is scored if the start of a cold wave is correctly predicted to within ± 1 day of the observed start date.

FIGURE 5 | Summary of hits, misses and false alarms for forecasts of cold wave events in Lesotho at lead times from 1 to 14 days in ECMWF control (ex-HRES) re-forecasts covering April–October 2004–2022. Note that the differences in the number of ‘true’ events (i.e., hits + misses) present at each lead time are due to the event-based approach combined with the irregular issue of re-forecasts; this behaviour is expected and also means that not all 40 events are represented at each lead time. Each ‘true’ event appears in approximately four re-forecasts (with different lead times). See Baugh et al. (2025) for further detail.

onset. The forecasts often predicted start dates that were wrong by only 1 day (Figure 4), with a deterioration in forecast performance at longer lead times. For example, of the 53 hits, 11 of these were at 1 day ahead and 8 at 2 days ahead, with the number of hits at lead times beyond 2 days ahead ranging between 1 and 5.

Relaxing the evaluation criteria to allow a hit for predictions within ± 1 day of the observed start date gave better performance. A summary of hits, misses and false alarms for this more flexible evaluation of the deterministic forecast is shown in Figure 5, where the POD exceeds 60% for most lead times up to 8 days ahead, with a false alarm rate of 43%–50% in the first 4 days of lead time, increasing beyond 4 days ahead.

Binary event forecasts were also derived from the ensemble probabilities by applying a range of decision thresholds (10%–100%), whereby forecasts were classified as predicting an event when the probability exceeded the chosen threshold. With a 4-day lead time, which is operationally critical for the EAP, these probability-based categorical forecasts achieved a threat score of around 50% (results not shown). Beyond 4 days, forecast skill declined, and threat scores fell below 30%, which, for the specific anticipatory actions and cost structures considered by users in this case study, suggests limited justification for triggering costly early actions at longer lead times. This conclusion was reached

through collaboration with the relevant decision-makers, and different users or decisions may place value on forecasts with substantially lower or higher skill. Overall, the evaluation highlighted a tendency of the ECMWF forecast, when combined with this specific threshold, to over-forecast the occurrence of cold wave events.

Development of the database of historical impactful events was viewed by users as highly valuable for refining thresholds and understanding impacts. User feedback also highlighted how metric preferences can vary between institutions: LMS expressed interest in seeing more traditional verification measures alongside the event-based evaluations that were favoured by LRCS. These insights reinforce the importance of tailoring evaluation and communication to user needs and capacities. Overall, the results provide recommendations for future updates of the EAP and an increased understanding of the forecast behaviour.

3.1.2 | Drought

The LRCS's EAP for drought is designed to enable anticipatory action based on forecasts of below-average rainfall, aiming to reduce humanitarian impacts such as food insecurity, crop loss and water scarcity. The EAP triggers are based on LMS's

seasonal precipitation forecast, which is derived from down-scaled global models and local observational knowledge and is recognised in the country for its accuracy and reliability.

The rainy season in Lesotho occurs from October to April and LMS issues its national seasonal forecast in September. Engagement with LRCS and LMS highlighted that evaluation of other seasonal forecasts would be beneficial to identify whether additional forecasts, available earlier, could provide valuable early guidance for more informal planning at LRCS ahead of the release of the official LMS forecast. The user-driven evaluation therefore focused on comparing global forecast products to understand which, if any, additional seasonal forecast may be most appropriate to use.

Four seasonal precipitation forecast products were evaluated: ECMWF's SEAS5 (Johnson et al. 2019), bias-adjusted SEAS5 and multi-model forecasts from the International Research Institute for Climate and Society (IRI) at Columbia University (IRI 2025), with the national forecasts from LMS as a benchmark. While the LMS forecast is available in September, the SEAS5 forecast is available from July and the IRI from August.

The EAP development process involved extensive consultation with stakeholders to incorporate local knowledge and professional expertise, drawing on the lived experiences of communities and insights of national and regional experts. This process identified a 25% forecast probability of precipitation in the December–January–February (DJF) season falling in the lower tercile of climatology as the threshold beyond which drought conditions are likely (this represents a relatively low probability threshold, indicating that even modest indications of below-average rainfall are considered sufficient to trigger concern). This threshold was defined through the EAP development process; the determination of this threshold was outside the scope of this study. It should be noted that such a threshold is context-specific and may not be optimal for all forecasting systems or decision contexts. Here, we adopt this existing user-defined EAP trigger to ensure operational relevance and evaluate how well this decision criterion is anticipated by different seasonal forecast products.

Evaluation was again based on a contingency table approach. In this case, forecasts exceeding the 25% probability threshold were compared against a binary record of historically documented drought events identified in literature over the period from 2001 to 2022. Forecasts produced in July, August and September were evaluated for their skill in predicting a drought, based on a threshold of a 25% forecast probability of precipitation in the DJF season falling in the lower tercile of climatology; not all forecast systems are available at all issue months. The comparison of forecast systems here is conditional on the fixed, user-defined decision threshold (25%) and reflects how well each system supports these specific decision criteria. It should not be interpreted as a general ranking of probabilistic forecast quality, as results may vary if different probability thresholds are considered.

Results indicated that none of the forecasts exhibited strong or consistently high skill across all metrics (POD, FAR and threat

score) and lead times. However, the ECMWF SEAS5 forecast (not bias-adjusted) performs better than the other forecasts at longer lead times (July) or similarly to the other forecasts in August and September (Figure 6). SEAS5's earlier availability is also operationally advantageous for early preparedness activities. The bias-adjusted SEAS5 forecasts require additional processing of the data before use but demonstrated no improvement over the original SEAS5, therefore providing no benefit over the publicly available original SEAS5 forecasts.

This user-driven evaluation was therefore able to recommend SEAS5 as the most appropriate forecast product for earlier preparedness in Lesotho, under the thresholds and decision context considered here, due to its reasonable forecast skill for the relevant event threshold, early release, open access and consistent availability. It should however be used in tandem with the forecast from LMS once this becomes available in September, providing added confidence given the noted trust in the national forecasts and the government's mandate to declare droughts.

3.2 | Precipitation Variability in Spain

The Living Lab in Los Pedroches—a rural area in northern Córdoba, south-central Spain, spanning 3612 km²—aimed to support agriculture and management of natural areas and water resources with information ranging from historical records to sub-seasonal forecasts and climate projections (Ropero Szymańska et al. 2025). Evaluation focused on the needs of olive tree growers rearing livestock (pigs) in the silvopastoral (integrated land management combining trees with pasture and livestock) system known as dehesa, and dairy farmers.

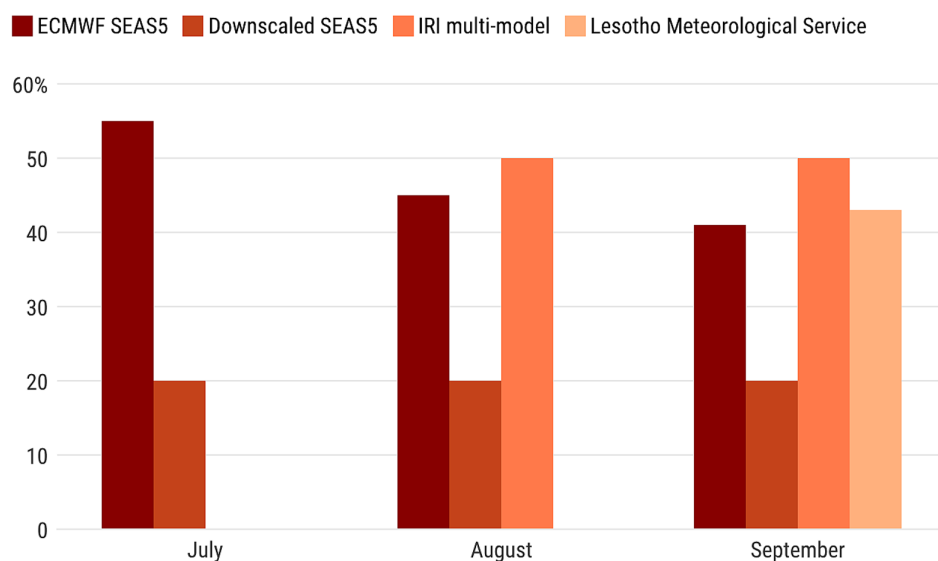
Decision timelines were created with the farmers to document the decisions they need to make and when, and the influence of the weather on them. Farmers typically indicated the 'ideal' amount of precipitation during key times of year. 'Adequate' ranges around these were co-defined for the purpose of seasonal forecast evaluation (Table 1). This evaluation tested the ability of two different seasonal forecasts to predict observed rainfall in 'adequate', 'less than adequate' and 'more than adequate' categories in October, autumn, winter and spring, using a contingency table approach.

ECMWF's SEAS5 forecasts (Johnson et al. 2019) and a version of SEAS5 post-processed using the HydroGFD3 reference dataset and the MIdAS bias adjustment methodology (Berg et al. 2021, 2022) were compared to rainfall observations covering the period 1993–2022 and provided by CREAM (Centre for Ecological Research and Forestry Applications), a public research centre in Spain.

For all available forecasts, produced on the first of each month from January 1993 to December 2022 and for the observations, the accumulated precipitation for each of October, autumn, winter and spring was calculated for grid squares covering the Los Pedroches region plus a 50 km buffer area and averaged over the region. Within the region used, observations were available for around 250 locations, and the observed data were interpolated to the same grid as the forecasts (0.25°).

Seasonal forecast evaluation for drought in Lesotho

Threat Score for prediction of a 25% probability of precipitation in the lower tercile of climatology in the December-January-February season, in four different forecasting systems providing forecasts in July, August and September



The threat score quantifies the proportion of observed or forecast events that were correctly predicted.

FIGURE 6 | Summary of forecast skill for the four evaluated seasonal precipitation forecasts (LMS, ECMWF SEAS5, ECMWF SEAS5 down-scaled and IRI multi-model), based on the threat score to quantify the proportion of observed or forecast events that were correctly predicted. A forecast is deemed to predict an event if it indicates at least a 25% probability of precipitation falling in the lower tercile of climatology in the December–January–February (DJF) season. Forecasts were evaluated for predictions of precipitation in the DJF season, for forecasts produced in September (all systems), August (SEAS5 and IRI) and July (SEAS5 only).

TABLE 1 | ‘Ideal’ and ‘adequate’ precipitation totals for each period of interest, as defined by farmers in the Los Pedroches region of Spain.

Time period	‘Ideal’ precipitation total (mm)	‘Adequate’ precipitation total (mm)
October	100	50–150
Autumn (SON)	150–200	100–300
Winter (DJF)	200–300	125–375
Spring (MAM)	200	100–300

Forecasts were compared to observations to assess the number of hits, misses, false alarms and correct negatives at a range of forecast probability thresholds from 10% to 100%. The maximum threat score reported in Table 2 therefore represents the highest value obtained across these thresholds, summarising the best event detection performance achieved across the range of probability thresholds considered. As such, it primarily reflects the discrimination ability of the forecast rather than its overall probabilistic performance.

Results showed that in general, performance was fairly poor for predictions of the ‘adequate’ precipitation ranges. This highlights the challenges associated with predicting specific and

relatively small ranges of precipitation for a specific location on seasonal timescales and also the tendency of seasonal forecasts to perform poorly when predicting more ‘normal’ conditions. For the ‘adequate’ and ‘less than adequate’ rainfall categories, maximum threat scores indicate that around half of events were correctly captured. The most promising forecast was for ‘less than adequate’ rainfall in winter, reaching a maximum threat score of 68% for the original SEAS5 forecasts, although skill is limited to the shortest lead times.

A higher threat score (82%) was obtained for the ‘adequate’ rainfall category in autumn; however, further analysis of hits, misses and false alarms shows no systematic improvement with decreasing lead time. This indicates that the high score arises from over-forecasting a frequently occurring category, rather than reflecting genuine predictive skill.

The ‘more than adequate’ category could not be assessed due to this occurring so infrequently that a robust evaluation was not possible.

Bias correction of the SEAS5 forecasts resulted in a small increase in forecast performance in five of the eight time period and category combinations, as the overall effect of the bias correction was to increase the total rainfall. However, these improvements in threat score are typically driven by just one less false alarm or one additional hit for a given category (Table 2).

TABLE 2 | Summary of maximum threat scores (the highest value obtained across forecast probability thresholds from 10 to 100%) and skill comparison for SEAS5 and bias-corrected SEAS5 forecasts of ‘less than adequate’ or ‘adequate’ precipitation totals in October, autumn, winter and spring in the Los Pedroches region of Spain, covering the period 1993 to 2020.

Summary of seasonal forecast skill for precipitation in Los Pedroches, Spain										
Maximum threat scores for prediction of ‘less than adequate’ and ‘adequate’ precipitation totals • ECMWF SEAS5 and bias-corrected SEAS5 forecasts (1993–2020)										
Time period	Precipitation category	Event count	Maximum threat score (SEAS5)	Probability threshold	Lead time (months)	Maximum threat score (bias-corrected)	Probability threshold	Lead time (months)	Better forecast	What led to the improvement?
October	Less than adequate	12	44%	40%, 70%	1	47%	50%, 40%	1, 2	Bias-corrected	1 false alarm
	Adequate	15	52%	20%	4, 6	54%	40%	2	Bias-corrected	1 false alarm
Autumn	Less than adequate	5	29%	40%	1	43%	60%	4	Bias-corrected	1 hit and 1 miss
	Adequate	23	82%	40%	4	82%	40%	4	Both	
Winter	Less than adequate	13	68%	40%	1	59%	30%	1	SEAS5	3 hits and 3 misses
	Adequate	13	50%	50%	1, 4	52%	60%	4	Bias-corrected	1 false alarm
Spring	Less than adequate	9	40%	30%	4	38%	20%	1	SEAS5	1 false alarm
	Adequate	20	66%	50%	2	69%	50%	2	Bias-corrected	1 hit and 1 miss

Note: Darker shading indicates better performance. The probability threshold(s) and lead time(s) (months) at which the maximum threat score was obtained are indicated.

For the case of ‘less than adequate’ rainfall in winter, the only category where SEAS5 showed promising forecast skill, the bias correction degraded the performance.

Future work could endeavour to assess seasonal forecasts from other sources and to consider a more ‘impact-based’ methodology, evaluating years which have historically been good, bad and ideal and relating these to the model climatology rather than focusing on specific observed precipitation totals. While the results in this case highlight challenges in the use of seasonal precipitation forecasts to predict specific ranges of precipitation, this remains important information for users and decision-makers to be aware of.

3.3 | Drought in the Netherlands

The Netherlands Living Lab is located in the Rijnland region, a small but highly managed sub-catchment at the downstream end of the Rhine delta. Around 72% of the area lies below sea level, and water is actively pumped out towards the Rhine, canals or the North Sea using an advanced drainage and discharge system. The Rijnland Water Authority is responsible for managing surface water quantity and quality, operating pumps, weirs and inlets to balance irrigation, flood protection and drinking water needs.

During dry periods, when there is a need for freshwater in the Rijnland region, this is sourced from the River Rhine, which enters the Netherlands at Lobith, making this a key reference point for national drought monitoring. Discharge here and commensurate water levels in the downstream delta of the Rhine influence navigation, water distribution and drought preparedness throughout the country. Forecasts of streamflow at Lobith are critical for water management in the Netherlands and are therefore the focus of user-driven evaluation in this living lab.

A seasonal streamflow forecasting system providing daily river discharge forecasts at Lobith out to 7 months ahead is evaluated.

Evaluation is performed against daily mean river discharge (m^3/s) at Lobith, with contingency table metrics computed for daily forecast values at each lead time. Lead time refers to the number of days or weeks between the forecast issue date and the target day (valid date) of the daily discharge forecast. The forecasts are issued once per month, initialised at the start of each calendar month and are generated using the E-HYPE (European Hydrological Predictions for the Environment) hydrological model of the Swedish Meteorological and Hydrological Institute (SMHI) (SMHI 2025), forced with bias-adjusted precipitation and temperature forecasts from ECMWF’s SEAS5 seasonal forecasting system (Johnson et al. 2019). Observations of river discharge at Lobith were used for comparison, with the evaluation covering the period 1994–2015. Further detail on the results summarised here can be found in Correa (2024).

Although both the forecasts and observations are available at daily resolution, the evaluation does not aim to assess the predictability of day-to-day variations in river discharge at seasonal lead times. Instead, the daily data are used to identify instances where flow falls below user-defined monthly thresholds, representing low-flow conditions relevant for drought management. An event is defined as a day on which the observed discharge falls below the relevant threshold and is forecast when the ensemble probability of falling below this threshold exceeds a given threshold, for example 30%. In this way, the evaluation focuses on the ability to discriminate between low-flow conditions and non-events over time, rather than to predict daily fluctuations.

Rijnland’s operational drought management decisions are based on monthly varying alert thresholds for low flows, which are defined from April to September. The thresholds are defined as 1000 m^3/s at Lobith in April and September, 1400 m^3/s in May, 1300 m^3/s in June, 1200 m^3/s in July and 1100 m^3/s in August (see Figure 8). Each of these thresholds was applied to forecasts predicting river flow throughout the period April–September from 1994 to 2015. A contingency table approach was taken to evaluate the ability of the forecasts to discriminate between low-flow events and non-events at different lead times and for different probability thresholds.

TABLE 3 | Contingency table indicating the number of hits, misses, false alarms and correct negatives at a 4-week lead time, based on seasonal forecasts of river flow from the E-HYPE (European Hydrological Predictions for the Environment) hydrological model forced with precipitation forecasts from ECMWF's SEAS5 seasonal forecasting system over the period 1994–2015, predicting a 30% probability of river flows falling below the 1400 m³/s threshold at the Lobith gauging station on the River Rhine in the Netherlands.

	Predicted	Not predicted
Observed	Hits: 21	Misses: 7
Not observed	False alarms: 34	Correct negatives: 70

At 2 weeks ahead, for example, with a probability threshold of 30%, the model correctly predicted 19 out of 22 observed days with flow below 1400 m³/s (86% POD), with 15 false alarms (14% false alarm rate). At 4 weeks ahead (Table 3), the number of correctly detected days with flow below 1400 m³/s was 21 out of 28 (75% POD) and the number of false alarms increased to 34 (33% false alarm rate). The difference in the number of observed events (days below the flow threshold) originates from the valid dates being different for a 2- and 4-week lead time for forecasts issued once a month. At longer lead times, POD decreased, and the false alarm rate increased, which may reduce user confidence in the forecast. For a well-calibrated forecast, lower probability thresholds will tend to increase the number of detected events, but at the cost of a higher false alarm rate, whereas higher probability thresholds will typically reduce false alarms while increasing the likelihood of missed events. The choice of probability threshold therefore reflects a trade-off that depends on the costs, benefits and risk tolerance associated with the decision being supported. This trade-off was presented to the users through the use of ROC curves (Figure 7), based on the results of the contingency table analysis.

The ROC curves were also used in co-exploring with users the most useful ways to present the evaluation results. For probabilistic forecasts, a set of contingency tables arises from varying aspects such as the forecast probability threshold, lead times and drought thresholds. ROC curves can provide a compact summary; however, users may prefer to see a set of contingency tables that more directly convey the outcomes relevant to their decisions across these dimensions. The users found that the contingency tables helped them to better understand forecast performance than the other metrics that were discussed, such as the ROC curves, and they appreciated the high detection rates for drought events. Contingency tables were found to offer more transparent information on the numbers of hits and false alarms, which are often important to users. These can be difficult to interpret from ROC curves. Alternative diagnostics could be explored in future work, such as relative economic value (Mylne 2002) (see Section 2.3.1) or precision-recall curves, which may provide additional insight for rare events by focusing on the balance between detected events and false alarms.

An additional approach used in this Living Lab was to co-evaluate, with representatives from the water authority, the forecasts for the drought event of 2003. During the co-evaluation

ROC curve for predicting low flow events in the Rhine at Lobith, The Netherlands

April–September • 4 week lead time • E-HYPE–SEAS5 • 1994–2015

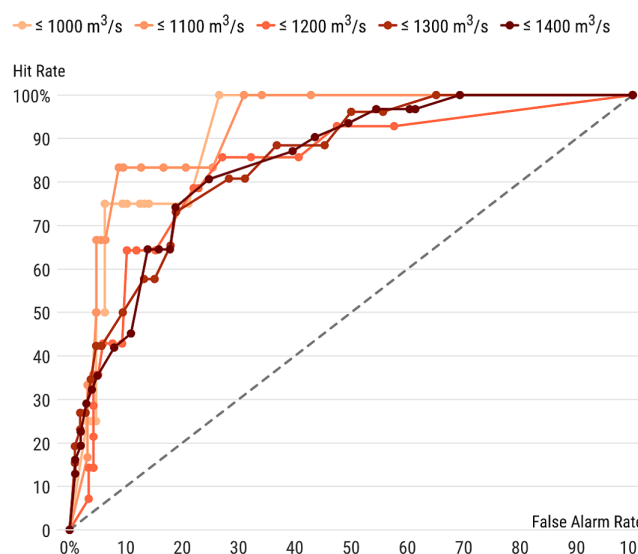


FIGURE 7 | Receiver operating characteristic (ROC) curves for predicting low flow events in the Rhine at Lobith, The Netherlands, for April–September with a 4-week lead time, based on the E-HYPE (European Hydrological Predictions for the Environment) hydrological model forced with precipitation forecasts from ECMWF's SEAS5 seasonal forecasting system over the period 1994–2015. Curves represent different river flow thresholds (≤ 1000 , ≤ 1100 , ≤ 1200 , ≤ 1300 and ≤ 1400 m³/s). The hit rate (true positive rate) is plotted against the false alarm rate (false positive rate), with higher predictive skill indicated by curves closer to the top-left corner of the diagram.

session, users reviewed forecast performance for this event by visually comparing the seasonal forecasts with the observed river discharge (example shown in Figure 8) to understand how well the forecasts captured reality. Users found this a useful exercise that would be beneficial to repeat for more events to build trust in the forecast performance. The example shown in Figure 8 highlights an important aspect of interpreting ensemble forecasts. In this case, the observed low flow conditions for July to October closely follow the lower bound of the ensemble distribution (the 10th percentile), indicating that the forecasting system was able to represent the extreme outcome within its spread. This illustrates that high-impact events may occur at the edges of the ensemble distribution, rather than near the ensemble mean, and as such, these low-probability outcomes should not be dismissed as unrealistic. Instead, they provide valuable information about plausible scenarios that may be critical for decision-making. In the context of this example, the use of a 30% probability threshold would mean that drought signals present in the extreme of the ensemble distribution are not captured as forecast 'events', despite indicating a risk of drought conditions, albeit at low probability. This highlights the importance of considering the full ensemble distribution, particularly for high-impact, low-probability events (Mylne 2026). The session also looked at more traditional verification metrics, such as the MAE. When analysing MAE, the users noted that forecast errors increased with lead time, with MAE values ranging from

Seasonal river flow forecast for the Rhine at Lobith

E-HYPE hydrological model with SEAS5 precipitation • April 2003 forecast

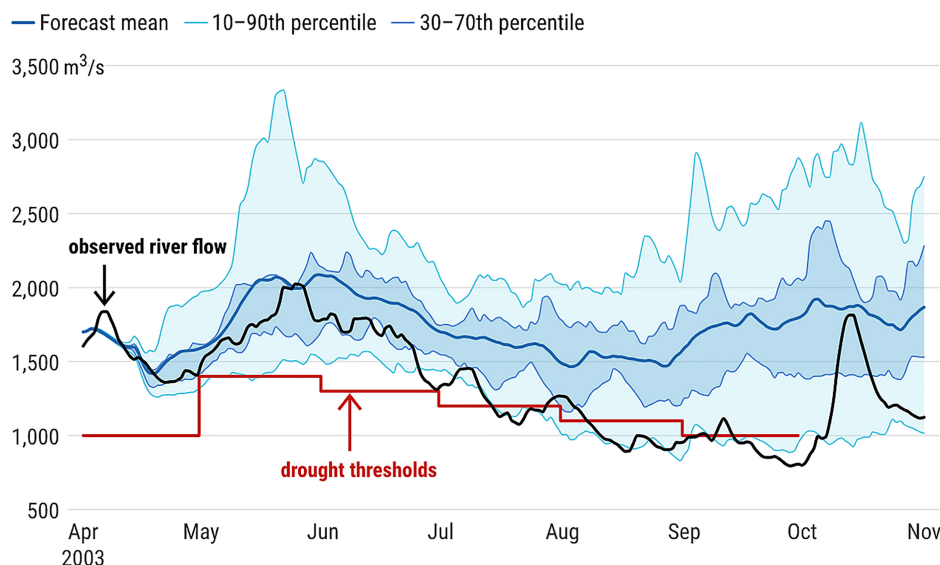


FIGURE 8 | Seasonal river flow forecast from 1 April 2003 from the E-HYPE (European Hydrological Predictions for the Environment) hydrological model, forced with bias-adjusted meteorological forecasts from ECMWF's SEAS5 seasonal forecasting system, ahead of the 2003 drought event. The blue line indicates the ensemble mean forecast, darker blue shading the 30–70th percentiles of the ensemble distribution and lighter blue shading the 10–90th percentiles. The black line shows the observed river flow throughout the period, for comparison with the forecasts, and the red lines show the monthly varying alert thresholds defined over the April–September drought monitoring season.

250 to $300 \text{ m}^3/\text{s}$ at 1–2 weeks lead time to around $500 \text{ m}^3/\text{s}$ at 6 weeks lead time.

Overall, the evaluation highlights that the forecasting system is able to reliably predict low-flow events. However, there is an indication of overprediction of events when compared to observations, based on the number of false alarms. Here, an important consideration is the probability threshold chosen and which threshold leads to the best balance of hits and false alarms for the decision-makers. A higher probability threshold would result in fewer false alarms. Based on this, users expressed that these seasonal forecasts could support drought alerts effectively in decision-making processes at lead times of 1–2 months.

Alongside the contingency tables, users explicitly requested more graphs of past events to allow comparison of forecasts and observations. They mentioned that these would also be useful to assess whether the forecasts are able to correctly capture increasing or decreasing trends in streamflow within the forecast horizon, even if there is a bias in the streamflow values themselves. This highlights the benefit of providing users with both robust statistics from a long-term evaluation and examples based on case studies of individual events for enhanced accessibility of the information.

3.4 | Low River Flows in Italy

In Italy, the focus of the Living Lab is on the prediction of low river flow conditions in the Secchia River, a tributary of the Po River in the Emilia-Romagna region of northern Italy. The area

is vulnerable to floods and droughts, with drought and associated water shortages resulting in water management challenges due to their impact on sectors such as agriculture, infrastructure, energy and tourism.

The interest for water management is in forecasting low flows in the next 6 months, so seasonal forecasts of river flow are evaluated. The forecasting system used is the E-HYPE hydrological model (SMHI 2025) driven with bias-adjusted meteorological forecasts from ECMWF's SEAS5 seasonal forecasts (Johnson et al. 2019) for the period 2003–2022, for which there is a record of river flow observations at the Lugo gauging station on the Secchia river, provided by the regional environment agency ARPAE Emilia-Romagna.

Low flow conditions are defined as those below a 'minimum environmental flow' (MEF) threshold that describes the minimum flow needed to sustain the ecology of the river. The MEF is typically defined as 10% of the mean annual observed river flow. At the Lugo gauging station, three thresholds are used to categorise the severity of low flows: green (river flow $> [\text{MEF} + 30\%]$), yellow ($[\text{MEF} - 30\%] < \text{river flow} < [\text{MEF} + 30\%]$) and red (river flow $< [\text{MEF} - 30\%]$). The MEF is $1.77 \text{ m}^3/\text{s}$. The upper bound of the yellow severity category (river flow $< [\text{MEF} + 30\%]$) is used here to evaluate the ability of the seasonal forecasts to predict low weekly-averaged river flows.

A contingency table approach was used to record the number of hits, misses and false alarms for each week of the forecast lead time from 1 to 30 weeks ahead. These numbers were evaluated for forecast probability thresholds from 10% to 100%.

Forecast evaluation for low river flow in the Secchia River, Italy

Threat score for seasonal forecasts of river flow below the yellow alert threshold at Lugo gauging station
2003–2022 • 1–30 weeks ahead • 10–100% forecast probability • E-HYPE with SEAS5

Threat score



Week	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1	7	7	7	7	7	2	3	3	3	0
2	15	12	5	5	5	3	3	0	0	0
3	16	10	6	7	2	2	0	0	0	0
4	16	9	7	5	0	0	0	0	0	0
5	18	13	5	5	5	2	0	0	0	0
6	15	6	7	2	0	0	0	0	0	0
7	16	8	4	2	2	2	0	0	0	0
8	17	14	11	4	4	0	0	0	0	0
9	20	10	10	5	3	0	0	0	0	0
10	18	10	3	0	0	0	0	0	0	0
11	10	6	7	5	2	0	0	0	0	0
12	13	8	4	4	0	0	0	0	0	0
13	17	9	5	3	0	0	0	0	0	0
14	15	9	0	0	0	0	0	0	0	0
15	12	4	2	0	0	0	0	0	0	0
16	14	7	0	0	0	0	0	0	0	0
17	14	2	0	0	0	0	0	0	0	0
18	16	4	0	0	0	0	0	0	0	0
19	12	2	0	0	0	0	0	0	0	0
20	11	7	0	0	0	0	0	0	0	0
21	21	6	0	0	0	0	0	0	0	0
22	16	4	0	0	0	0	0	0	0	0
23	19	5	0	0	0	0	0	0	0	0
24	20	4	0	0	0	0	0	0	0	0
25	17	2	0	0	0	0	0	0	0	0
26	16	7	0	0	0	0	0	0	0	0
27	15	0	0	0	0	0	0	0	0	0
28	13	0	0	0	0	0	0	0	0	0
29	12	2	0	0	0	0	0	0	0	0
30	13	2	0	0	0	0	0	0	0	0

The threat score quantifies the proportion of observed or forecast events that were correctly predicted.

FIGURE 9 | Legend on next page.

FIGURE 9 | Summary of forecast skill for seasonal river flow forecasts based on the E-HYPE hydrological model forced with bias-adjusted SEAS5 meteorological forecasts for the Lugo gauging station on the Secchia River, Italy. Based on the threat score, quantifying the proportion (%) of observed or forecast events that were correctly predicted. An event is defined as weekly-averaged river flow falling below the yellow alert threshold (flows < [minimum environmental flow + 30%]). Threat scores provided for lead times of 1–30 weeks ahead, at forecast probability thresholds from 10% to 100%. Evaluation period 2003–2022. Evaluated using river flow observations at the Lugo gauging station.

Results indicate limited skill across all lead times and probabilities (Figure 9). Beyond 14 weeks ahead, the threat score was 0 for all probability thresholds higher than 20%. The threat scores reached a maximum of 21% for a forecast of a 10% probability of flows below the yellow threshold. The forecasts fail to predict (miss) many low-flow events. The 10% probability threshold gives the largest number of hits but is accompanied by a larger number of false alarms. One reason the forecast probability threshold is low (10%) is due to a positive bias (over-prediction of river flows) in the forecasts during the July to October low flow season.

The relatively narrow range of threat scores (10%–21%) across the range of lead times at the 10% probability threshold and the lack of systematic decline with increasing lead time suggest limited predictive skill beyond that expected from climatological frequencies and a lack of sensitivity to forecast horizon, consistent with weak underlying predictability. The shortest lead time (1 week) also shows lower threat scores (7%) than subsequent lead times for 10%–30% probability thresholds. This may reflect sampling variability or the influence of biases that are more pronounced at shorter lead times and further emphasises the limited and inconsistent skill for the decision-relevant thresholds considered here.

An alternative method of evaluation is to derive thresholds from the historical forecasts (ECMWF 2025b), which circumvents this issue of model biases relative to observations. By defining thresholds using the forecast climatology rather than observations, events are evaluated relative to the model's own distribution, reducing the influence of systematic biases and isolating the model's ability to discriminate between events and non-events. Using this approach, the forecasts showed slightly better skill; however, the threat scores were still very low, with a maximum value of 25% at 3 weeks ahead. For the most extreme category of low flows (< 10% of the historical climatology for the time of year), the maximum threat score was just 11% at 1 week ahead.

To put these results in context, the 2022 drought event was investigated. River flow at the Lugo gauging station consistently fell below the yellow threshold and crossed the red threshold from July to October 2022. The forecasts predicted flows to remain within the normal range during this period. Although earlier forecasts from March and April did predict slightly lower-than-average flows, the subsequent forecasts from May onwards did not and so the forecasts would not have provided sufficient information about the potential severity and duration of the drought for this event.

Given the low scores, it is not possible to conclude that these seasonal forecasts would provide useful information for anticipation of low river flows in the Secchia River. The flow in the catchment is largely driven by rainfall, which is difficult to predict beyond medium-range (typically up to 2 weeks ahead)

timescales. Seasonal river flow forecasts can be more skilful in catchments with a large upstream drainage area, large contributions from snowmelt or large contributions from groundwater—none of these are the case in the Secchia, making seasonal forecasts challenging.

To better understand the sources of forecast error, the river flows resulting from driving the E-HYPE model with meteorological data from the ERA5 reanalysis dataset were also evaluated. As ERA5 represents the best available estimate of past weather (although it is not a perfect representation of reality, containing its own uncertainties in precipitation estimates), this provides an indication of how well the hydrological model performs when the meteorological forcing is based on reality rather than a forecast scenario. The results from a traditional verification indicated a moderate performance, based on a KGE score of 0.51 (where 1 would represent a perfect match, values around 0 indicate performance comparable to using the mean observed flow as a predictor and values in the range 0.4–0.6 can be considered to indicate moderate skill). This indicates that uncertainty also arises from the rainfall-runoff modelling itself in this catchment, in addition to the challenges of predicting precipitation at seasonal timescales. This result highlights the importance of considering multiple sources of uncertainty in seasonal river flow forecasts, particularly in catchments where predictability is inherently limited. It also emphasises the potential usefulness of combining user-driven evaluation, considering the decision-making context, with traditional verification to help diagnose the results.

3.5 | Drought in Georgia

Seasonal forecasts of low river flows are also the focus of the Living Lab in Georgia, in the Alazani and Iori river basins. These basins are in the eastern part of Georgia, where cold winters and hot, dry summers result in pronounced seasonal variations in river flow. Runoff dynamics are driven by snowmelt in spring and convective rainfall in summer, contributing to complex low-flow conditions during the dry season and presenting challenges for sectors such as agriculture, water supply and ecosystems, particularly under increasing pressure from climate change and water demand from irrigation and hydropower sectors.

A variety of decision-makers rely on seasonal forecasting in the Alazani–Iori basin, with different decision-making contexts. For example, water managers and hydropower operators are interested in river flow and volumes entering the Sioni Reservoir on the Iori River, particularly during summer, while farmers and environmental managers are concerned about lower water availability, also during summer. The National Environmental Agency (NEA) plays a central role in monitoring river flows, issuing hydrological bulletins and disseminating early warnings to this range of stakeholders.

Verification metrics for seasonal river flow forecasts for Shaqriani, Alazani-Iori basin, Georgia

European Flood Awareness System (EFAS) • 1991–2018

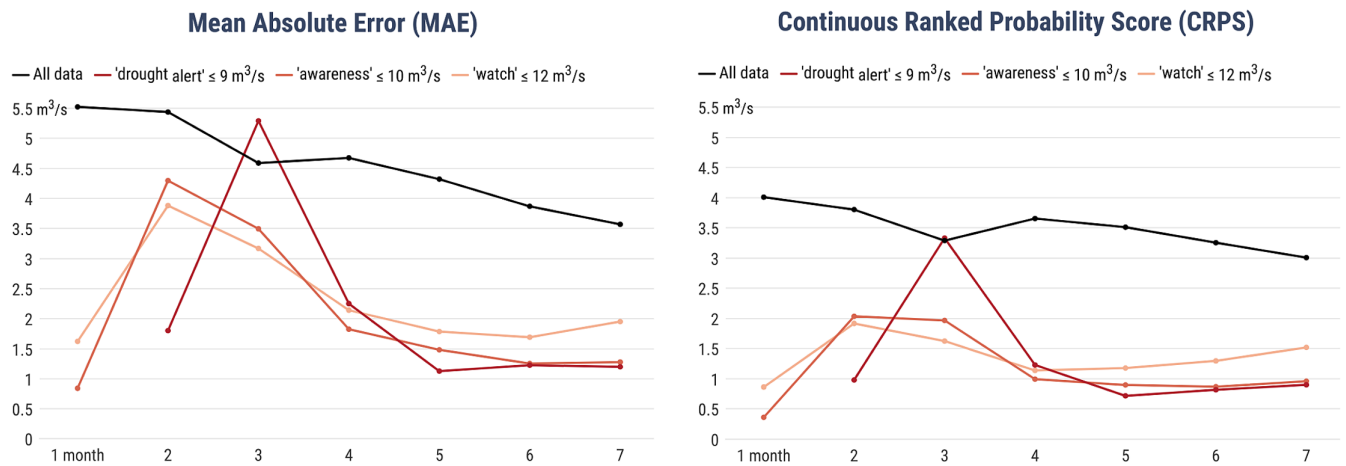


FIGURE 10 | Mean absolute error (MAE, left) and mean Continuous Ranked Probability Score (CRPS, right) of seasonal forecasts of river flow from the European Flood Awareness System (EFAS) at lead times of 1–7 months at the Shaqriani gauging station in the Alazani–Iori basin in Georgia. Black lines indicate the performance for all data; red for river flows $\leq 9 \text{ m}^3/\text{s}$ ('drought alert' threshold), dark orange for river flows $\leq 10 \text{ m}^3/\text{s}$ ('drought awareness' threshold) and light orange for river flows $\leq 12 \text{ m}^3/\text{s}$ ('drought watch' threshold). For the threshold-specific lines, MAE and CRPS are computed conditionally by restricting the analysis to days on which the observed river discharge falls below the corresponding low-flow threshold, therefore representing conditional error metrics for drought-relevant conditions for comparison to overall forecast performance.

Through co-evaluation workshops with the NEA and the use of decision timelines, a simple alert system using statistical drought thresholds based on observed river flows was developed, as no formal alert thresholds had previously been established. These thresholds were then used in the evaluation of the forecasts. The Shaqriani gauging station in the Alazani River was chosen as the location for evaluation due to its long record of observations. A daily river flow below $12 \text{ m}^3/\text{s}$ indicates a drought 'watch', below $10 \text{ m}^3/\text{s}$ 'attention' and at or below $9 \text{ m}^3/\text{s}$ an 'alert'. These thresholds are based on the 33rd, 20th and 10th percentiles of the observed record. These percentiles were also calculated from the seasonal streamflow forecasts as the forecast climatology to assess potential model bias.

The forecasts evaluated are the seasonal streamflow forecasts from the Copernicus Emergency Management Service (CEMS) European Flood Awareness System (EFAS; Copernicus 2025a, 2025b) from 1991 to 2018. The evaluation again focuses on a contingency table approach, supported by traditional verification metrics. The latter, based on MAE and CRPS for all forecasts initiated from April to September (Figure 10), indicates higher forecast accuracy for lower flows. They also indicate a decrease in error at longer lead times, which differs from the typical expectation that forecast error increases with lead time. In this case, smaller errors at longer lead times may reflect a combination of factors, including contraction of forecast variance towards climatological mean conditions at longer lead times and a reduction in systematic biases at short lead times. As lead time increases, ensemble forecasts tend to lose sharpness and converge towards climatological conditions, which can reduce average error metrics even as decision-relevant information is diminished. This highlights a key limitation of aggregated error-based metrics such as MAE and CRPS, as lower scores at longer lead times do not necessarily imply greater usefulness for decision-making.

The contingency table approach (Table 4) concluded that the forecast quality decreases with increasing lead time for all three thresholds. At 1 month ahead, using a forecast probability threshold of 40% for the 'drought watch' threshold, the contingency table shows two hits, no false alarms and three missed events. This indicates that while some drought conditions are detected, some events also remain unanticipated at this threshold and lead time. By 3 months ahead, there is a noticeable drop in performance. Some false alarms are introduced, with 5 false alarms compared to 4 misses and 1 hit, and at 6 months ahead the uncertainty is large with a substantial increase in false alarms (46, compared to 4 hits and 1 miss).

A co-evaluation exercise with stakeholders included looking at forecasts for a particular drought event in 2018 compared to observed data. It was noticed that observed flows were higher than model flows from May to July, likely caused by the model failing to capture late snowmelt and/or convective rainfall events. This indicates a model bias towards higher flows than observed and is consistent with the contingency table results (Table 4), which show an increase in false alarms and a decrease in correct negatives at longer lead times. Varying the probability threshold at different lead times may give better results and could be explored in future. From August to December, during a period of lower flows, the model and observations aligned more closely. This exercise proved useful to supplement the long-term evaluation and provide additional insight into the information behind the evaluation, and highlights the value of combining traditional verification metrics with user-driven evaluation. The decrease in MAE and CRPS at longer lead times (Figure 10), together with the contingency table results (Table 4), suggests that the forecasts tend to converge towards climatological conditions. While this reduces average error metrics, the underlying bias leads to an overprediction of low-flow events, demonstrating how

TABLE 4 | Contingency table indicating the number of hits, misses, false alarms and correct negatives for seasonal forecasts of river flow from the European Flood Awareness System (EFAS) over the period 1991–2018, predicting a 40% probability of river flows falling below the user-defined ‘drought watch’ threshold ($\leq 12 \text{ m}^3/\text{s}$) at the Shaqriani gauging station in the Alazani–Iori basin in Georgia.

Lead time	Event occurrence	Predicted	Not predicted
1 month ahead	Observed	Hits: 2	Misses: 3
	Not observed	False alarms: 0	Correct negatives: 74
3 months ahead	Observed	Hits: 1	Misses: 4
	Not observed	False alarms: 5	Correct negatives: 71
6 months ahead	Observed	Hits: 4	Misses: 1
	Not observed	False alarms: 46	Correct negatives: 30

traditional metrics can help to diagnose and explain behaviour seen in the contingency table-based evaluation.

Overall, the case study exercise combined with the contingency table results were found by the decision-makers to be useful, informative and relevant to the decision-making context. The MAE was also found to be useful, with local forecasters considering the results to be within an acceptable range, but the CRPS, alongside ROC curves associated with the contingency table analysis, was found to be challenging to understand and apply to the decision-making context. This highlights the importance of using evaluation metrics that are technically robust but accessible and useful to users. Further detail on the results summarised here can be found in Correa (2024).

4 | Discussion

4.1 | Common Themes

Several common themes emerged from the case studies, despite the diverse decision-making contexts in each of the Living Labs. First, contingency tables proved to be a highly effective way of evaluating forecasts in a decision-relevant way, presenting results and engaging with decision-makers. Their outputs—hits, misses, false alarms and correct negatives—are intuitive and map directly onto real-world decisions and events. Users generally found these results easier to understand than more abstract evaluation metrics and scores, and results highlighted that the way in which the results are visualised is critical. Nevertheless, the apparent simplicity of contingency tables can mask methodological choices that can have a substantial influence on results, such as how to define the start of an event, how to handle multi-day events or whether to allow a tolerance in timing. Careful co-definition of these choices with decision-makers and transparent documentation are therefore essential. Further to this, decision-making criteria provide an easy entry point for conducting user-driven evaluation, particularly if these criteria are quantitative. A key aspect of decision-relevant evaluation is understanding not only the thresholds that define an event, but also the probability thresholds at which users choose to act, which reflect their tolerance for risk and the relative cost of missed events and false alarms. However, if these are not already clearly defined, it can be challenging to initially identify the best thresholds to use. This may require an iterative and participatory process to

identify thresholds or criteria, and the results of the evaluation may also help to further refine thresholds used in decision-making. The emphasis on understanding and managing the trade-off between missed events and false alarms is consistent with decision-theoretic perspectives on probabilistic forecasting, in which forecast probabilities only become meaningful through their translation into action thresholds (Mylne 2002).

Event-based evaluation was often, but not always, preferred over more traditional evaluation, as it aligns more closely with operational decisions triggered by discrete thresholds. Additionally, in some cases a relatively small number of events can result in the need to combine event-based approaches with other longer-term analyses. Co-evaluation of individual historical events was also found to be particularly valuable, as this provided opportunities for dialogue and reflection and helped to contextualise long-term statistics and build trust in the understanding of the forecasts and their evaluation. These case studies must be carefully chosen to ensure they are representative; however, it must be clear that the evolution of individual cases may differ from one event to the next, so choosing one historical event may not reflect the typical skill of a forecasting system.

Traditional evaluation metrics, in some cases, provided useful additional diagnostic insight, for example, by explaining false alarms through a systematic bias, but these metrics were often less accessible to users. ROC curves and CRPS in particular risk sending overly optimistic messages if used in isolation, particularly for rare events, where correct negatives dominate the scores. However, when interpreted appropriately, they can provide useful information for comparing forecasts, for example, in terms of overall probabilistic performance (CRPS) or discrimination ability (ROC). Nevertheless, these measures do not directly indicate the usefulness of forecasts for specific decisions, which depend on user-defined thresholds and actions. Future work could consider the use of precision-recall curves (Saito and Rehmsmeier 2015) or the threshold-weighted CRPS (twCRPS; Gneiting and Ranjan 2011). Institutions also differed in their preferences, with some valuing the familiarity of traditional metrics, while others focused on the key messages provided by the counts of hits, misses and false alarms. This highlights that while traditional metrics remain important for scientific model development and research, they should not be relied upon as the sole evidence of decision relevance; flexibility and tailoring of

the approach to the decision context and user perspective are essential.

Finally, the evaluations underscored that while in some cases forecasts can be useful at predicting certain types of events relevant for certain decisions, in other cases forecasts can have very limited skill at thresholds most relevant to users. The case study of low river flows in Italy illustrates that while a forecast may demonstrate scientific skill, for example when assessed relative to a climatology or using traditional verification metrics, it may still fail to provide actionable decision-support skill for a given decision context. In the Secchia catchment, the seasonal forecasts showed limited capability to predict low-flow conditions at the thresholds and lead times of operational relevance, reflecting the inherent predictability constraints at seasonal timescales. Identifying such situations where statistical skill may be present but where forecast uncertainty, biases or lead-time limitations prevent reliable discrimination between actionable and non-actionable conditions is a critical outcome of user-centred evaluation. Doing so may prevent overconfidence or inappropriate use of forecasts and supports more transparent dialogue between forecast providers and users about what information can and cannot be relied upon. While this can be a challenging result to convey, it reinforces the importance of transparency and underscores the important role of purveyors/intermediaries in creating a shared understanding between forecast producers and users, including when communicating poor skill and the limitations of certain methodologies or metrics alongside their benefits. Users will be quick to identify shortcomings in the forecasts once they begin to rely on forecasts for decision-making, and clear communication of both strengths and weaknesses is essential to maintain trust and allow for effective decision-making.

4.2 | Advantages and Challenges of the Framework

This user-driven evaluation framework proved flexible and adaptable to diverse decision-making contexts for a range of event types in different countries and involving different stakeholders. Its strengths lie in linking evaluation to concrete decision thresholds, in promoting transparency and in fostering dialogue between forecast providers, intermediaries and users.

The iterative process of understanding the decision context, defining thresholds and co-examining results fostered mutual learning, with some cases showing that evaluation itself can help to further refine and improve the decision thresholds to tailor them to the forecasts being used. However, challenges remain. Data availability was a limiting factor in some Living L4abs, particularly where observational records were sparse or where a clear link between impacts and thresholds was not well established. Communicating results, particularly those for probabilistic forecasts, in user-friendly formats remains an area for further development. While contingency tables provide an accessible entry point, they still require careful explanation and visualisation choices can strongly influence how results are interpreted.

The framework currently focuses on the evaluation of forecasts and climate services, but further work should see the findings more systematically fed back into service design and improvement, consistently working towards improved forecasts and climate services. Overall, the framework has demonstrated its value as a structured yet adaptable approach, but it should be seen as iterative and evolving, with continued work towards refining methods for defining and evaluating events in ways that are both scientifically robust and relevant for users.

4.3 | A Holistic View of Forecast Evaluation

A comprehensive evaluation of forecasts and climate services requires the integration of both traditional verification and user-driven approaches and evaluation of the salience and legitimacy of the service. Traditional statistical metrics provide measures of accuracy and are critical for model development and benchmarking performance over time and can be used to help interpret the user-driven scores and diagnose problems. However, they should be complemented by approaches that assess how forecasts perform in decision-making contexts, including whether they meet user-defined, decision-relevant thresholds and support real-world choices (Pagano et al. 2025). To do this, it is important to consider users' local knowledge, capacities and preferences and allow them to guide the choice of evaluation measures and metrics. Where explicit cost-loss relationships can be defined, value-based approaches such as relative economic value provide a coherent decision-theoretic framework for comparing forecasts and selecting probability thresholds for specific decisions. Understanding forecast performance from the user's perspective is equally as important as statistical accuracy, as relevance, trust and usability often determine whether information leads to meaningful decisions. How information is visualised and communicated also significantly affects its interpretation and use; evaluation should therefore also seek to consider whether forecasts are presented in formats that users can readily understand and apply. These views are in line with the WMO Global Framework for Climate Services (GFCS; gfcs.wmo.int) that emphasises the need for evaluations that consider not only scientific quality but also the salience and legitimacy of climate services in real-world contexts.

5 | Conclusions and Recommendations

This paper has presented a framework for user-centred, decision-relevant evaluation of forecasts and climate services, supported by examples of its application in several diverse case study regions across Europe and Africa. The findings align with established forecast-value frameworks showing that decision relevance emerges from the interaction between forecasts, user thresholds and decisions rather than from forecast accuracy alone (Richardson 2000; Mylne 2002). Based on the insights and results from these case studies, evaluating forecasts using a user-centred, decision-relevant perspective, we conclude and recommend the following:

The actions users intend to take following a forecast or derived alert and the decision process towards these actions have been used to ensure user relevance of the forecast evaluation.

Participatory methods such as the decision timeline exercise were effective for climate service co-creation and for user-centred forecast evaluation. Understanding what the user needs to decide and do based on a forecast and why helps to shape our understanding of the decision-making context and which aspects of the forecast or climate service are most important to evaluate. This emphasis on starting from decision-making needs is not only important for evaluation but also for how forecasts are selected and used more broadly. Recent work on ensemble forecast applications similarly highlights the importance of defining use cases from the perspective of decisions and actions (e.g., Mylne et al. 2025).

Recommendation: Take the time to understand the decisions that need to be made, including why (what actions need to be taken based on the decision?), when (such as timing during the year) and how (based on what specific information/forecast/thresholds) they are made.

While four of the five Living Labs were interested in forecasting drought conditions, the local definition of drought was different for each—three focused on low river flows, although with varying threshold definitions and one on seasonal precipitation variability. This further highlights that working directly with decision-makers and taking the first step of understanding the decision-making context and their specific challenge is important in order to design an effective user-driven forecast evaluation. This includes, for example, specifying variables and event alert thresholds. It also highlights that thresholds can be refined for use with a specific forecast or for a certain type of event, and these thresholds themselves can sometimes be further refined using evaluation of the forecasts.

In some of the case studies, the event definitions (thresholds) were indeed further specified or refined over time. This was typically for one of four reasons: (1) one meeting may not provide enough discussion time to cover all aspects in detail; (2) sometimes time and further discussion is needed to develop a common understanding between the decision-makers and researchers (3) sometimes; following time for reflection, decision-makers were able to define their challenges and operations more precisely and provide us with more detail and (4) in one example (cold waves in Lesotho), the user-driven evaluation itself produced results that helped to further refine and improve the thresholds at which action is taken based on a forecast, due to improved understanding of the forecast performance.

Recommendation: Include multiple iterations and opportunities for discussion when defining events and alert thresholds.

Identifying user-preferred evaluation approaches resulted in new insights. Despite discussions and workshops dedicated to understanding as best as possible the decision-making context and challenges to develop an informed evaluation approach, in some cases, presenting a range of metrics and results to decision-makers led to opportunities to gain additional perspectives on how these are used and understood. For example, while the ROC curve can be used to concisely summarise a range of results based on a contingency table approach, it was found to be too complex to use effectively. In reality, while it could seem more

complex to provide a larger set of contingency tables, this way of presenting the information was much preferred and easier to understand and offers transparency around the actual number of correctly predicted events, misses and false alarms. This also highlights a broader challenge, as while more rigorous verification metrics can provide valuable diagnostic insight, they are not always readily interpretable in decision-making contexts without careful translation and communication. Bridging this gap between statistical robustness and practical usability remains a key area for further development. Discussing evaluation approaches can also reveal the forecast characteristics that the decision-makers value the most. For example, in one case the ability of a forecast to correctly predict increasing or decreasing trends during the forecast horizon was of more concern than a bias in the values predicted. In another, the start date of an event was found to be a critical piece of information over other aspects of the forecast.

Recommendation: The evaluation approach/metric and visualisation you prefer and most easily understand may not be the most accessible to each decision-maker. Listen to and iterate with decision-makers on the approach most suitable for them and be open to presenting the findings in different ways depending on who will be using them. This also requires further development of approaches, in collaboration with users, to communicate rigorous verification methods in ways that remain scientifically robust and easily interpretable across a range of decision contexts.

The co-creation and user-centred evaluation process undertaken through the I-CISK project has provided information on the performance of forecasts and climate services that has proven to be useful and usable for the decision-makers we collaborated with. Overall, users reported that the results of the user-driven evaluation were more intuitive and understandable than the results from traditional verification. It also provided valuable experiences and insights for those of us undertaking the evaluation.

Future work may also endeavour to co-evaluate other aspects of wider climate services, going beyond forecast skill to understand the salience, legitimacy and credibility of the service, including aspects such as the delivery mechanism, visualisation, interpretation or decision support. In the co-creation of tailored climate services, these aspects should be considered holistically alongside the forecast skill, and opportunities may also arise to directly integrate forecast skill information into climate services, for example, through skill dashboards or direct integration into forecast visualisations, further improving the potential usability.

Recommendation: Our strongest recommendation is that the forecasting and climate services communities—from operational centres and service providers to researchers and advisory bodies—commit to user-centred evaluation of forecasts and climate services and ensure that the results are made publicly, freely and easily accessible to those who depend on them for critical decisions.

With an increase in user-centred evaluation, forecasts and climate services will become more useful, usable, understood and ultimately more used to inform decisions.

Author Contributions

Rebecca Emerton: conceptualization, investigation, writing – original draft, methodology, visualization, writing – review and editing. **Daniele Castellana:** data curation, investigation, writing – review and editing. **Deborah Dotta Correa:** data curation, investigation, writing – review and editing. **Marc van den Homberg:** methodology, investigation, writing – review and editing. **Micha Werner:** conceptualization, investigation, writing – review and editing. **Katherine Egan:** methodology, data curation, investigation, writing – original draft. **Yiheng Du:** data curation, writing – review and editing. **Calum Baugh:** methodology, data curation, investigation, writing – original draft. **Sumiran Rastogi:** methodology, writing – review and editing. **Ilias Pechlivanidis:** methodology, data curation, writing – review and editing. **Schalk Jan van Andel:** conceptualization, methodology, investigation, writing – original draft. **Emie Klein Holkenborg:** data curation, investigation, writing – review and editing.

Acknowledgements

This research was funded by the European Union's Horizon 2020 Research and Innovation Programme under the Grant Agreement Number 101037293: Innovating Climate services through Integration of Scientific and local Knowledge (ICISK). Further information and resources associated with this project can be found at www.icisk.eu. The input from the I-CISK Multi-Actor Platform members within the Living Labs, such as providing data, local knowledge and insights into decision-making processes, is gratefully acknowledged.

Funding

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme (101037293).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

This study makes use of data from a range of sources. Seasonal meteorological forecasts from ECMWF (SEAS5), seasonal river flow (re-)forecasts from the European Flood Awareness System (EFAS) and the ERA5-Land reanalysis dataset are freely and openly available via the Copernicus Climate Data Store (<https://cds.climate.copernicus.eu>). ECMWF medium-range forecast data can be accessed through ECMWF's Meteorological Archival and Retrieval System (MARS), subject to licensing. Other data were provided by national services and stakeholders under research collaborations and as such are not publicly available. Please contact the authors for further information.

References

- Anticipation Hub. 2024. "EAP: Lesotho—Cold Wave." <https://www.anticipation-hub.org/fr/telecharger/file-5041>.
- Anticipation Hub. 2025. "Anticipatory Action in 2024: A Global Overview." <https://www.anticipation-hub.org/download/file-4974>.
- Bamzai-Dodson, A., and R. A. McPherson. 2022. "When Do Climate Services Achieve Societal Impact? Evaluations of Actionable Climate Adaptation Science." *Sustainability* 14, no. 21: 14026. <https://doi.org/10.3390/su142114026>.
- Baugh, C., A. Egan, K. Klein, et al. 2025. *Assessment of Existing and Tailored Climate Services Using a Range of User-Driven Evaluation Metrics*. I-CISK Deliverable 3.4. <https://icisk.eu/resources/>.
- Berg, P., F. Almén, and D. Bozhinova. 2021. "HydroGDF3.0: A 25 km Global Precipitation and Temperature Data Set Update in Near-Real Time." *Earth System Science Data* 13, no. 4: 1531–1545. <https://doi.org/10.5194/essd-13-1531-2021>.
- Berg, P., T. Bosshard, W. Yang, and K. Zimmermann. 2022. "MidASv0.2.1 0 Multi-Scale Bias Adjustment." *Geoscientific Model Development* 15, no. 15: 6165–6180. <https://doi.org/10.5194/gmd-15-6165-2022>.
- Calvel, A., M. Werner, M. van den Homberg, et al. 2020. "Communication Structures and Decision Making Cues and Criteria to Support Effective Drought Warning in Central Malawi." *Frontiers in Climate* 2: 578327. <https://doi.org/10.3389/fclim.2020.578327>.
- Cantone, C., H. Ivars Grape, S. El Habash, and I. G. Pechlivanidis. 2023. "A Co-Generation Success Story: Improving Drinking Water Management Through Hydro-Climate Services." *Climate Services* 31: 100399. <https://doi.org/10.1016/j.cliser.2023.100399>.
- Copernicus. 2025a. "European Flood Awareness System." <https://european-flood.emergency.copernicus.eu/en>.
- Copernicus. 2025b. "Seasonal and Sub-Seasonal Hydrological Outlooks." <https://european-flood.emergency.copernicus.eu/en/seasonal-outlook>.
- Correa, D. D. 2024. "Co-Evaluation of Climate Services: Integrating Forecaster- and User-Oriented Perspectives for Seasonal Streamflow Forecast Evaluation." Master's thesis, IHE Delft Institute for Water Education.
- Crochemore, L., S. Materia, E. Delpiazzi, et al. 2024. "A Framework for Joint Verification and Evaluation of Seasonal Climate Services Across Socioeconomic Sectors." *Bulletin of the American Meteorological Society* 105: E1218–E1236. <https://doi.org/10.1175/BAMS-D-23-0026.1>.
- de Perez, E. C., B. van den Hurk, M. K. van Aalst, B. Jongman, T. Klose, and P. Suarez. 2015. "Forecast-Based Financing: An Approach for Catalyzing Humanitarian Action Based on Extreme Weather and Climate Forecasts." *Natural Hazards and Earth System Sciences* 15: 895–904. <https://doi.org/10.5194/nhess-15-895-2015>.
- Ebert, E., B. Brown, M. Goerber, et al. 2018. "The WMO Challenge to Develop and Demonstrate the Best New User-Oriented Verification Metric." *Metallurgische Zeitschrift* 27: 435–440.
- ECMWF. 2024. "Implementation of IFS Cycle 48r1." <https://confluence.ecmwf.int/display/FCST/Implementation+of+IFS+Cycle+48r1>.
- ECMWF. 2025a. "Integrated Forecasting System." <https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model>.
- ECMWF. 2025b. "EFAS Sub-Seasonal Outlook. Copernicus Emergency Management Service—Floods User Documentation." <https://confluence.ecmwf.int/display/CEMS/EFAS+Sub-seasonal+outlook>.
- European Commission: Directorate-General for Research and Innovation, R. Street, M. Parry, et al. 2015. *A European Research and Innovation Roadmap for Climate Services*. Publications Office of the European Union. <https://data.europa.eu/doi/10.2777/702151>.
- Gneiting, T., and R. Ranjan. 2011. "Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules." *Journal of Business and Economic Statistics* 29, no. 3: 411–422. <https://doi.org/10.1198/jbes.2010.08110>.
- Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez. 2009. "Decomposition of the Mean Squared Error and NSE Performance Criteria: Implications for Improving Hydrological Modelling." *Journal of Hydrology* 377: 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>.
- Harrowsmith, M., M. Nielsen, C. Jaime, et al. 2020. "Impact-Based Forecasting for Early Action, the Future of Forecasts." <https://doi.org/10.13140/RG.2.2.12366.89920>.
- IRI. 2025. "Seasonal Climate Forecasts." <https://iri.columbia.edu/our-expertise/climate/forecasts/seasonal-climate-forecasts/>.
- Johnson, S. J., T. N. Stockdale, L. Ferranti, et al. 2019. "SEAS5: The New ECMWF Seasonal Forecast System." *Geoscientific Model Development* 12: 1087–1117. <https://doi.org/10.5194/gmd-12-1087-2019>.

- Lemos, M. C., C. J. Kirchhoff, and V. Ramprasad. 2012. "Narrowing the Climate Information Usability Gap." *Nature Climate Change* 2: 789–794. <https://doi.org/10.1038/nclimate1614>.
- Lemos, M. C., and B. J. Morehouse. 2005. "The Co-Production of Science and Policy in Integrated Climate Assessments." *Global Environmental Change* 15: 57–68. <https://doi.org/10.1016/j.gloenvcha.2004.09.004>.
- Masih, I., N. Van Cauwenbergh, A. Ziogas, et al. 2022. *D1.1 Characterization of the I-CISK Living Labs*. I-CISK Project. <https://doi.org/10.5281/zenodo.17583446>.
- Matheson, J. E., and R. L. Winkler. 1976. "Scoring Rules for Continuous Probability Distributions." *Management Science* 22, no. 10: 1087–1096. <http://www.jstor.org/stable/2629907>.
- McClure, A., J. Daron, S. Bharwani, et al. 2024. "Principles for Co-Producing Climate Services: Practical Insights From FRACTAL." *Climate Services* 34: 100492. <https://doi.org/10.1016/j.cliser.2024.100492>.
- Muñoz-Sabater, J., E. Dutra, A. Agustí-Panareda, et al. 2021. "ERA5-Land: A State-of-the-Art Global Reanalysis Dataset for Land Applications." *Earth System Science Data* 13, no. 9: 4349–4383. <https://doi.org/10.5194/essd-13-4349-2021>.
- Murphy, A. H. 1993. "What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting." *Weather and Forecasting* 8: 281–293. [https://doi.org/10.1175/1520-0434\(1993\)008<0281:WIAGFA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2).
- Mylne, K. 2002. "Decision-Making From Probability Forecasts Based on Forecast Value." *Meteorological Applications* 9, no. 3: 307–315. <https://doi.org/10.1017/S1350482702003043>.
- Mylne, K. 2026. "Probability Forecasts—Part 2: Will People Understand?" *Weather* 81, no. 4: 100–106. <https://doi.org/10.1002/wea.70016>.
- Mylne, K., N. Roberts, D. Walters, et al. 2025. "Classification of Use Cases for Ensemble Weather Forecasts." *Bulletin of the American Meteorological Society* 106: E1981–E1999. <https://doi.org/10.1175/BAMS-D-24-0183.1>.
- Nyamakura, B., I. Masih, M. Werner, L. Hermans, and G. Jewitts. 2025. "Typologies of Climate Service Co-Creation Approaches in Practice." *Climate Services* 40: 100607. <https://doi.org/10.1016/j.cliser.2025.100607>.
- Pagano, T. C., B. Casati, S. Landman, et al. 2024. "Challenges of Operational Weather Forecast Verification and Evaluation." *Bulletin of the American Meteorological Society* 105, no. 4: E789–E802. <https://doi.org/10.1175/BAMS-D-22-0257.1>.
- Pagano, T. C., E. E. Ebert, and M. Khanarmuei. 2025. "Enhancing Forecast Verification in National Meteorological and Hydrological Services." *Meteorological Applications* 32: e70051. <https://doi.org/10.1002/met.70051>.
- Pappenberger, F., M. H. Ramos, H. L. Cloke, et al. 2015. "How Do I Know if My Forecasts Are Better? Using Benchmarks in Hydrological Ensemble Prediction." *Journal of Hydrology* 522: 697–713. <https://doi.org/10.1016/j.jhydrol.2015.01.024>.
- Pechlivanidis, I. G., Y. Du, J. Bennett, et al. 2025. "Enhancing Research-to-Operations in Hydrological Forecasting: Innovations Across Scales and Horizons." *Bulletin of the American Meteorological Society* 106: E894–E919. <https://doi.org/10.1175/BAMS-D-24-0322.1>.
- Richardson, D. S. 2000. "Skill and Relative Economic Value of the ECMWF Ensemble Prediction System." *Quarterly Journal of the Royal Meteorological Society* 126, no. 563: 649–667. <https://doi.org/10.1002/qj.49712656313>.
- Ropero Szymańska, N., N. Hernández-Mora, and L. De Stefano. 2025. "Characterizing Adaptation Responses to Drought Risk of Livestock Farmers in the Spanish Dehesa Agroforestry System." *Frontiers in Environmental Science* 13: 1540818. <https://doi.org/10.3389/fenvs.2025.1540818>.
- Saito, T., and M. Rehmsmeier. 2015. "The Precision-Recall Plot Is More Informative Than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PLoS One* 10, no. 3: e0118432. <https://doi.org/10.1371/journal.pone.0118432>.
- SMHI. 2025. "About the HYPE Model." <https://hypeweb.smhi.se/about-us/about-the-model/>.
- van den Homberg, M., S. Rastogi, L. De Stefano, et al. 2024. *D2.5 User-Centred Validation of Climate Risk Knowledge Integration: Using Decision Timelines for Collecting, Understanding, and Integrating Local Knowledge*. I-CISK Project. <https://doi.org/10.5281/zenodo.17588901>.
- Vaughan, C., and S. Dessai. 2014. "Climate Services for Society: Origins, Institutional Arrangements, and Design Elements for an Evaluation Framework." *WIREs Climate Change* 5: 587–603. <https://doi.org/10.1002/wcc.290>.
- Wens, M., M. Schaafsma, M. Werner, et al. 2025. *D2.8 A Guideline for End-User Centred Co-Creation of Climate Services Across Europe and Beyond*. I-CISK Project. <https://doi.org/10.5281/zenodo.17589096>.
- White, C., C. J. White, D. I. V. Domeisen, et al. 2022. "Advances in the Application and Utility of Subseasonal-to-Seasonal Predictions." *Bulletin of the American Meteorological Society* 103: E1448–E1472. <https://doi.org/10.1175/BAMS-D-20-0224.1>.
- Wilks, D. S. 2011. *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Academic Press.
- WMO. 2025. *Guidelines on the Verification of Hydrological Forecasts*. WMO. <https://library.wmo.int/idurl/4/69478>.