A framework for joint verification and evaluation of seasonal climate services across socio-economic sectors

Louise Crochemore,*[a,l] Stefano Materia,[b,m] Elisa Delpiazzo,[c] Stefano Bagli,[d] Andrea Borrelli,[b] Francesco Bosello,[c] Eva Contreras,[e] Francesco Dalla Valle,[g] Silvio Gualdi,[b] Javier Herrero,[e] Francesca Larosa,[h,j] Rafael Lopez,[i] Valerio Luzzi,[d] Paolo Mazzoli,[d] Andrea Montani,[f,k] Isabel Moreno,[i] Valentina Pavan,[f] Ilias Pechlivanidis,[a] Fausto Tomei,[f] Giulia Villani,[f] Christiana Photiadou,[a,n] María José Polo,[e] Jaroslav Mysiak[h]

[a] *Swedish Meteorological and Hydrological Institute, Hydrology Research Unit, 60380 Norrköping, Sweden*

[b] *CSP Division, Centro Euro-Mediterraneo sui Cambiamenti Climatici, 40127 Bologna, Italy*

[c] *ECIP Division, Centro Euro-Mediterraneo sui Cambiamenti Climatici, 30175, Marghera, Italy; Department of Environmental Sciences, Informatics and Statistics Department, Ca' Foscari University of Venice, 30172, Mestre, Italy; European Institute for the Economy and Environment, 30175, Marghera, Italy*

[d] *GECOsistema Srl, 47521 Cesena, Italy*

[e] *Fluvial Dynamics and Hydrology Research Group, Andalusian Institute for Earth System Research, University of Córdoba, 14071 Córdoba, Spain*

[f] *Arpae—Regional Agency for Prevention, Environment and Energy of Emilia-Romagna, 40122 Bologna, Italy*

[g] *Enel Green Power, MTS - Dams & Civil Infrastructures Safety - Hydrology and Hydraulic Analysis, Mestre 30172, Italy*

[h] *RAAS Division,Centro Euro-Mediterraneo sui Cambiamenti Climatici, 30175, Marghera, Italy; Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, 30172, Mestre, Italy*

[i] *Physics for Renewable Energies Research Group, Universidad de Córdoba, 14071 Córdoba, Spain*

[j] *KTH Climate Action Centre, Royal Institute of Technology, Stockholm, Sweden*

[k] *European Center for Medium-Range Weather Forecasts, Shinfield Rd, Reading RG2 9AX, United Kingdom*

[l] *Univ. Grenoble Alpes, CNRS, INRAE, IRD, Grenoble INP, IGE, 38000 Grenoble, France*

[m] *Barcelona Supercomputing Center, Barcelona, Spain*

[n] *European Environment Agency, Kongens Nytorv 6, 1050 Copenhagen, Denmark*

Corresponding author: Louise Crochemore, louise.crochemore@univ-grenoble-alpes.fr

Early Online Release: This preliminary version has been accepted for publication in *Bulletin of the American Meteorological Society*, may be fully cited, and has been assigned DOI 10.1175/BAMS-D-23-0026.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

Assessing the information provided by co-produced climate services is a timely challenge given the continuously evolving scientific knowledge and its increasing translation to address societal needs. Here we propose a joint evaluation and verification framework to assess prototype services that provide seasonal forecast information based on the experience from the H2020 CLARA project. The quality and value of the forecasts generated by CLARA services were firstly assessed for five climate services utilizing the Copernicus Climate Change Service seasonal forecasts and responding to knowledge needs from the water resources management, agriculture, and energy production sectors. This joint forecast verification and service evaluation highlights various skills and values across physical variables, services and sectors, as well as a need to brigde the gap between verification and user-oriented evaluation. We provide lessons learnt based on the service developers' and users' experience, and recommendations to consortia that may want to deploy such verification and evaluation exercises. Lastly, we formalize a framework for joint verification and evaluation in service development, following a transdisciplinary (from data purveyors to service users) and interdisciplinary chain (climate, hydrology, economics, decision analysis).

SIGNIFICANCE STATEMENT

Tools to communicate climate-related information to users, typically dam managers, irrigation consortia, or energy producers are fast evolving to answer societal needs. It is crucial to estimate the quality of the provided information, along with economic, environmental and/or societal gains. Here we exemplify how to assess information quality and potential gains in five services that provide data and information for hydropower, solar power, irrigation and water reservoirs in Europe and South America. Based on this work, we recommend: (1) for service developers to well anticipate such quality and value assessments, due to the number of actors to be involved, (2) for flexibility when screening how to quantify quality and gain to account for decision contexts and (3) for sustained funding or collaborating platforms to ensure the iterative co-evaluation process.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

CAPSULE

A transdisciplinary framework to verify seasonal forecasts and evaluate climate services allows assessing the value of forecast-based services for real-life decision-making and provides a cross-sector perspective.

1. Introduction

Climate and water services convey climate-relevant knowledge and information (processed, structured and well-communicated data interpreted for decision-making contexts) for horizons ranging from weeks to decades ahead. Climate services, hereafter used to refer to both climate and water services, have been recognized as pivotal instruments to support decision-making at all levels, from public entities to business operators (Alexander and Dessai, 2019; Boon et al. 2022; Brasseur and Gallardo, 2016) and across sectors, such as disaster risk reduction, energy production, water resources management or agriculture (e.g. Street et al., 2019, Troccoli et al., 2018).

One key conclusion that emerged from previous service generation initiatives is the prominent role of users to ensure actionable information, and thus the need for co-production that ensures continuous interactions between service users, developers and data purveyors (Bremer et al., 2019; Photiadou et al., 2020; Cantone et al., 2023). Co-development and co-evaluation are participatory steps of co-production (together with co-design and co-delivery; Mauser et al., 2013) which focus on developing a climate information chain addressing user needs and on assessing the value of climate services, respectively. By building trust through direct participation, co-development and co-evaluation promote the sustainable use of climate services, and contribute to the common reflection on how climate data may address user needs (Clements et al., 2013).

The need for "*standards for what constitutes 'quality' climate information*" and improved skill indicators to support users with measurable quality criteria has been previously highlighted (Vaughan and Dessai, 2014). Forecast verification, which assesses the quality of forecasts in reproducing retrospective outcomes (Jolliffe and Stephenson, 2003; Wilks, 2011), is somewhat common practice, though not systematic, in openly available climate services. Yet, forecast verification does not account for the decision being informed by forecasts, and therefore does not capture the benefit (utility) and economic value of the forecast information (Giuliani et al., 2020). The literature on socio-economic forecast evaluation has thus grown, bridging the gap between the climate and economical science communities (e.g. Katz and Lazo,

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

2011; Laugesen et al., 2022). Nevertheless, examples of the value (economic, societal or environmental) attained by climate services in use and how these assessments are practically carried out are far less common (Vaughan et al., 2019).

This paper aims, in a first stage, to present the skill and value assessments of five service prototypes based on seasonal forecasts (Section 2). These services were developed within the Horizon 2020 innovation action CLARA project (Climate forecasts enabled knowledge services, https://www.clara-project.eu, 2017-2020). The project aimed to support the co-production of climate service prototypes across sectors, and demonstrate the implementation of the open-access and large-scale Copernicus Climate Change Service (C3S). It involved service developers, sectorial users, hydro-climate scientists, and economists in co-producing climate services and in ensuring their longevity through sustainable business models. Quantitative outcomes from the verification and evaluation exercises performed within CLARA are presented (Sections 3 and 4) and qualitative outcomes from developers' and users' viewpoints are discussed (Section 5). We draw recomendations for verifiaction and evaluation exercises based on the CLARA experience (Section 6) and formalize a framework for assessing the quality and value of climate services providing information at the seasonal time scale and focusing on operational decision-making (Section 7), before concluding (Section 8).

2. Design of the verification and evaluation exercises within CLARA

2.a. Climate services

The services studied here were selected among the 14 co-produced within CLARA, because they make use of the seasonal forecasts available from C3S (rather than decadal information or climate projections). They aim to support decision-making through interactive forecast visualization and provide sector-oriented forecasts targeting water resource management, agriculture, hydropower production and solar energy production, and potential users such as water authorities, reservoir managers, and energy producers. Table 1 presents the objectives, developers and users of each service, and Fig. 1 shows the data and information flowchart of each service.

The way seasonal forecasts are generated in each service, depends on the specificities of each natural system and on the expertise of the service providers (Fig. 1 and Table 2a). For instance, the SCHT service (Smart Climate Hydropower Tool) was developed using Artificial Intelligence (AI) to relate meteorological predictands with reservoir inflows, and SEAP (Solar Energy Assessment and Planning Tool) is based on a statistical model that predicts photovoltaic

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

plant production based on solar, wind and meteorological conditions. ROAT (Reservoir Operation Assessment Tool) and SHYMAT (Small Hydropower Management and Assessment Tool) offer tailored information at local scale through bias adjustment and downscaling of available C3S seasonal forecasts. WRI (Water Resources for Irrigation) is based on a mechanistic impact modeling meant to translate hydro-meteorological information into variables of interest for the agriculture sector. The concepts behind these services are transposable to different geographical areas. Nevertheless, the verification and evaluation presented hereafter focus on their original target areas in Europe (i.e. Spain and Italy) and South America (i.e. Colombia).

All services rely on multiple data sources, including observations acquired locally in real-time or simulations from impact models, and meteorological and hydrological forecasts from C3S available through the Climate Data Store (CDS) (see details in Table 2a): SEAS5 from the European Center for Medium-range Weather Forecasts (ECMWF, Johnson et al., 2019), SPS3 from the Euro-Mediterranean Center on Climate Change (CMCC, Sanna et al., 2016), System 7 from Météo-France (Batté et al., 2019) and GFCS 2.0 from DWD (Baher et al., 2015). Seasonal hydrological forecasts were made available to service developers from the C3S proof-of-concept service for the European water resources management sector (https://climate.copernicus.eu/operational-service-water-sector). These forecasts are based on the E-HYPE hydrological model from the Swedish Meteorological and Hydrological Institute (SMHI) and the ECMWF SEAS5 meteorological forecasts (Pechlivanidis et al., 2020).

Table 1. General information on the climate services detailing their sector of application, their name and a corresponding reference, their aim, target audience, developer, user within CLARA and a link for access.

Sector	**Service name (developer)**	**Aim**	**Target audience**	**Developer**	**User involved in the co-development**	**Access**
Water resources management	ROAT - *Reservoir Operation Assessment Tool*	Support operations of multi-objective reservoirs by anticipating (i) drought risks and setting a "scarcity level", (ii) water excess from snowmelt to avoid damages downstream the dam.	Water authorities, reservoir managers	University of Córdoba	Technicians of the Béznar-Rules reservoir	https://www.uco.es/dfh/ROAT/
Agriculture	WRI - *Water Resources for*	Support decision-making for both water procurement and water allocation by	Irrigation water management authorities (e.g. water	Arpea	Land Reclamation and Irrigation	https://servizi-

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

	Irrigation (Villani et al., 2021)	anticipating water irrigation needs to increase efficiency and reduce irrigation water and energy consumption	procurement and allocation agencies)		Consortium of Romagna and Burana	gis.arpae.it/moses/home/index.html
Hydropower production	SCHT - *Smart Climate Hydropower Tool* (Essenfelder et al. 2020)	Simplify decision-making processes by predicting hydropower production for management and financial planning	Hydropower energy producers	GECOSistema	Enel Green Power	https://gecosistema.com/scht
	SHYMAT - *Small Hydropower Management and Assessment Tool* (Contreras et al. (2020a; 2020b))	Support operation of run-of-river plants such as (i) planning the operation and maintenance tasks, (ii) anticipating energy production, and (iii) setting the turbine level depending on river flow	Technicians in charge of the run-of-river plants and managers of hydropower companies. Energy market operators, river basin authorities and consultants	University of Córdoba	Endesa technicians	https://www.uco.es/dfh/SHYMAT/
Solar energy production	SEAP - *Solar Energy Assessment and Planning Tool*	Automate both spatial and operational assessment of utility-scale photovoltaic power plants by providing information about the optimal tracking system for dynamic collectors.	Photovoltaic plant managers	University of Córdoba	Magtel technicians	https://www.uco.es/investigacion/proyectos/seap/

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

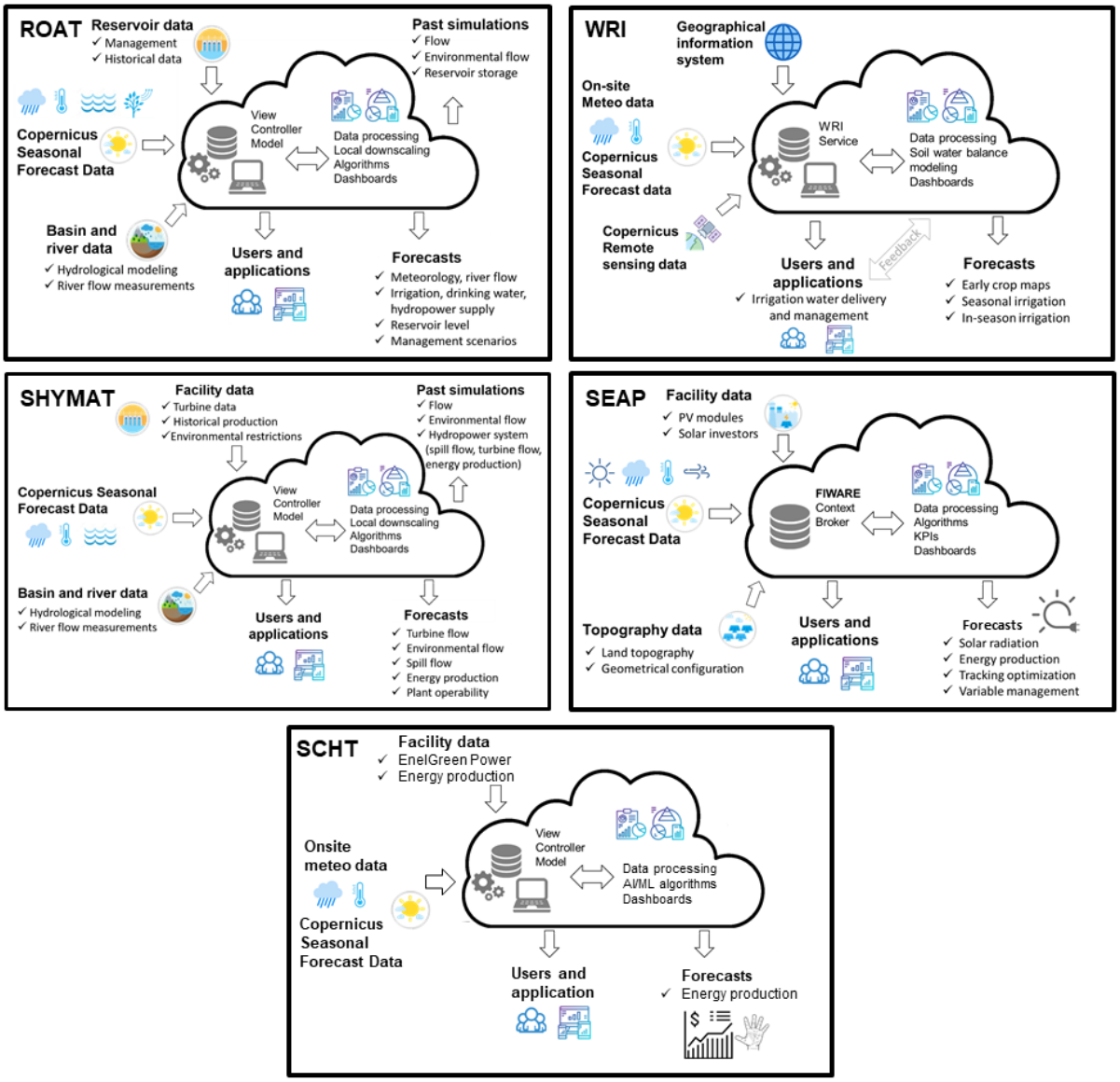

Fig. 1. Workflow diagrams of the climate services. Each service workflow displays the service input data, including the seasonal forecast data, the methods applied within the service (such as pre-processing, impact models, user interface) and the type of output data. A detailed description of the verified and evaluated components is available in Table 2.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

Table 2. Service characteristics during the CLARA exercise (a), and parameters of the joint verification (b) and evaluation (c) exercises for each of the services.

a)

Service	**ROAT**	**WRI**		**SCHT**		**SHYMAT**	**SEAP**
Area considered	Órgiva River Basin in Sierra Nevada (495 km^2), Spain	Irrigated area of the Irrigation Consortium of Romagna (930 km^2), Italy	Irrigated area of the Irrigation Consortium of Burana (600 km^2), Italy	Betania hydropower plant, Magdalena River (13,000 km^2), Colombia	Guavio hydropower plant on the Guavio River (1,500 km^2), Colombia	Water-contributing area to Pampaneira hydropower plant (82.5 km^2)	"El Molino", solar installation, located in Cordoba, Spain
Service output forecasts	Dynamical probabilistic forecasts bias adjusted with quantile mapping	Hybrid probabilistic forecasts from a weather generator supplying precipitation and temperature		Artificial Intelligence-based statistical deterministic forecasts		Dynamical probabilistic forecasts downscaled for the system location	Statistical deterministic forecasts
Copernicus input forecasts	Forecasts of river flow as inflows to the reservoir (source: ECMWF/SMHI)	Forecasts of precipitation and temperature (source: CMCC/ECMWF /DWD/MétéoFrance - 100 members)		Forecasts of precipitation and temperature (source: CMCC)		Forecasts of river flow as inflows to the hydropower plant (source: ECMWF/SMHI)	Forecasts of wind, radiation, precipitation and temperature (source: CMCC)

b)

Service	**ROAT**	**WRI**	**SCHT**		**SHYMAT**	**SEAP**
Quality-assessed variable	Bias adjusted river flow (m^3/s)	Crop irrigation water needs (m^3/ha)	Inflow to the reservoirs (Mm^3/month)		Bias adjusted river flow (m^3/s)	Photovoltaic production (kWh/month)
Assessment period	22 years (1989/01-2014/12 excluding 1990, 1997-1999)	6 years (2011-2016)	26.5 years (1993/01-2019/08)		35 years (1981/01-2015/11)	10 years (1997/01-2006/12)
Time step	Monthly	3-month	6-month	3-month	Monthly	Monthly
Issue dates and horizons	Monthly for each of the 7 months ahead - *12x7 forecast values per year*	Once a year for the summed volume over June, July, August - *1 forecast value per year*	Monthly for the inflows averaged over the next 6 months - *12 forecast values per year*	Monthly for the inflows averaged over the next 3 months - *12 forecast values per year*	Monthly for each of the 7 months ahead - *12x7 forecast values per year*	Monthly for the month ahead - *12 forecast values per year*

Reference - *truth*	River flow observations from Hidrosur and Béznar-Rules Reservoir System	Crop water needs reanalysis data based on ERA5-land (one grid cell (0.1°x0.1°))	Inflow observations from Enel Green Power	River flow observations from Endesa	Photovoltaic production reanalysis data based on ERA5
Benchmark – *reference forecast*	Ensemble forecast based on historical flows from the previous 10 years	Forecast based on ERA5-land climatology (1993-2010) as input to the model	Enel linear model fed with Copernicus seasonal forecasts (CMCC)	Ensemble forecast based on historical flows from the previous 10 years	Forecast based on ERA5 reanalysis as input to the photovoltaic production model

c)

Service	**ROAT**	**SCHT**	**SHYMAT**	**SEAP**
Variable used to define the states of the world (indicator)	Stored water volumes (hm^3) - *derived from river flow and reservoir rules*	Stored water volume (Mm^3) - *derived from reservoir inflows and reservoir rules*	River flow (m^3/s)	Daily irradiation on photovoltaic modules based on clearness index *kt* (*kt* values from 0 to 0.88) - *correlated with photovoltaic production*
Test period	From 2010 to 2015	From 2011 to 2014	1-year (2010/2011) and 4-year (2011-2015) periods	2013
Time scale	Monthly	Monthly	Monthly	Daily
States of the world	5 states: (i) flood control; (ii) normal state; (iii) early warning; (iv) alert; (v) emergency	3 states: water volume (i) inside (ii) below (iii) above the normal range defined by the management rule in the previous 10 years	4 states: flow (i) higher than maximum flow; (ii) inside maximum range; (iii) inside minimum range; (iv) lower than minimum flow	8 states based on 8 different levels of *k*t
Alternative knowledge source	Historical data of stored water volume and inflows	Climatological mean of water volumes over the last 30 years	Experience-based knowledge on river flow thresholds	Actual collectors' tracking method
Payoff	Payoff based on availability of water for different uses	Payoff based on produced energy	Payoff based on produced energy	Payoff based on solar radiation (proxy for profits)
Unit of measure for service value	% variation in performance index based on final users' judgements (0-10)	% variation in performance index based on final users' judgements (0-10)	% variation in performance index based on final users' judgements (0-10)	kWh/m^2

2.b. Methodology

Due to the transdisciplinary and time-limited nature of the exercise, compromises were made in selecting verification and evaluation methodologies. The metrics used to convey the forecast quality and service value were chosen so as to be easily computable by the service developers within the time of the project, and easily understandable and translatable to users. This led, in some cases, to the simplification of known forecast attributes (reliability) and in the selection by users of reduced but contrasting sets of situations and actions. Neither the verification nor the evaluation metrics used here are part of the recommendations. As discussed thereafter, verification metrics should be tailored for each climate service and decision context, highlighting forecast attributes relevant to the evaluation.

Hydro-meteorological observations were used as *reference* (or truth) in the verification and evaluation of the services, i.e. river flow observations (Table 2b; SHYMAT, SCHT and ROAT). When local observations were not available, ERA5 (ECMWF Reanalysis v5) Land meteorological reanalysis data (Hersbach et al., 2020) forced the impact models used in the services to generate *reference* simulations (e.g. temperature and precipitation in the case of WRI to simulate crop water needs; solar radiation, temperature and wind in the case of SEAP to simulate photovoltaic production).

2.b.1) Designing the verification exercise

Verification was performed for each service by comparing hindcasts (retrospective forecasts) of the service with a *reference* (or truth) over a long enough time period with respect to different *forecast attributes*. The relative quality of the service with respect to a reference forecast (or *benchmark)* was then assessed with *skill* scores (see Table 2b). The assessment was tailored to the specificities of each service, meaning that the seasonal forecast set, target variable, geographical area, and verified period and forecast attributes vary with the service based on user needs and data availability (Table 2b). Note that verification should ideally be performed with at least 30 years of hindcasts. The short verification periods for WRI and SEAP thus lead to significant uncertainties in their verification thereafter.

Even though probabilistic forecasts are recommended to convey the known uncertainties at seasonal lead times, not all decision frameworks incorporate probabilistic information yet. Some services thus provided deterministic forecasts as a support to current decision flows. For the services involving probabilistic forecasts (ROAT, SHYMAT), an

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

indicator of *reliability* was complemented with an indicator of *sharpness* (often assessed alongside reliability; Gneiting et al., 2007). For the services based on deterministic forecasts (SCHT and SEAP), *correlation* and *accuracy* were assessed. In the case of WRI, an event-based verification assesses *accuracy* in identifying above- and below-average situations. All attributes were assessed for each initialisation month and forecast month relevant to the service, using cross-validation when parameters were adjusted (see Appendix A for metrics formulations). *Skill scores* assess the relative quality with respect to a forecast *benchmark* (Hargreaves, 2010; Appendix A), which consists of forecasts that decision-makers would use in the absence of the service (Table 2b). The benchmark choice was tailored to each service, with the goal to represent a realistic alternative to the proposed service. For four services, forecasts were compared to a climatological benchmark: ROAT and SHYMAT verified their forecasts against observed climatology, while WRI and SEAP used model climatology (i.e. simulations generated by forcing impact models with historically observed meteorological fields) due to the lack of observations of crop water needs and photovoltaic production. For SCHT, the forecasts are AI-based and were compared to a regression model also forced with Copernicus meteorological forecasts. *Skill scores* were computed for each forecast attribute, with positive values indicating that the service outperforms the benchmark.

2.b.2) Designing the evaluation exercise

Service *value* (Table 2c) stems from the benefits from forecast-based decisions. Co-evaluation can be ex-ante (during co-development) demonstrating potential values, or ex-post (after co-development) demonstrating real-life service value after co-delivery. Here, ex-ante service value was estimated for all services but WRI due to limited in-situ observations related to irrigation practices.

Many reviews on climate service evaluation methods (e.g. Vaughan et al., 2019; Clements et al., 2013) have shown the impossibility to define a one-fits-all approach. The methodology chosen here is a Bayesian probabilistic framework (Murphy ,1993; Katz and Murphy, 1997; Katz and Lazo, 2011) with the advantage to produce a quantitative assessment fostering the climate service adoption, especially if the user is private and profit-seeking. The challenge of this fairly simple methodology resides in harmonizing inputs from different climate services and decision-making contexts, so that they are usable for the methodology and meaningful for the users (Appendix B describes how it was practically performed). In CLARA, this approach was complemented by a qualitative evaluation (not presented here) to assess

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

users' perception of the services beyond just profits. That (ex-post) evaluation could however only be performed once the services were developed.

The link between the forecasts and the operative decisions is represented through an *indicator* considered by the user when making decisions. It could be the variable forecast by the service or a derived variable. Some critical indicator *thresholds* delimit intervals named *states of the world*. The users provide information about the *payoffs* they can gain when they decide to act based on a state of the world. An important aspect of the assessment is the metrics used to represent payoffs. A monetary or physical metric is easily understandable by users and the general public, though sometimes not possible because the service has not yet been used in a real-life decision-making context or the payoff information is sensitive. Alternatively, performance can be measured by asking decision-makers to assign scores to combinations of forecast states and actions (see "Unit of measure" in Table 2c).

The *expected value* is assessed for different knowledge sources based on the payoff of actions in each state of the world, the capacity of the knowledge source to forecast this state of the world, and on the probability of occurrence of this state of the world (for a comprehensive review and a practical example see Delpiazzo et al., 2022). First, the expected value of *perfect information* (Pope et al., 2017) is derived from the effective realizations (based on reference observations, or simulations) representing a climate service that perfectly forecasts the states of the world. It is instrumental as it demonstrates the maximum potential value of a service. The *expected value* of an *alternative knowledge* source (based on business-as-usual practices) is then assessed. In the case of SCHT and ROAT, the alternative knowledge is climatology-based, for SHYMAT, it is the experience-based performance (replicating the same actions of a past period), and in SEAP, the operators' original tracking methods was used as reference. Lastly, once available, the expected value of the *climate service knowledge* is assessed.

The outcomes of the value assessment are the *effective value of the service* and its *maximum potential value*. The former is the difference in expected value between the climate service in its given development phase and the alternative knowledge (Table 2c). The latter derives from the difference between the expected value of perfect information when past observations are used as proxies for a climate service that always forecasts correctly and the alternative knowledge.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

3. Outcomes of forecast verification

When verifying, deterministic forecasts of water volumes (SCHT) and photovoltaïc production (SEAP) show the highest accuracy, while probabilistic forecasts of crop water needs (WRI) show the lowest accuracy (not shown). Such results could be expected due to the high predictability of river flows as opposed to the low predictability of precipitation, that weigh directly on the performance of the water needs forecasts. In fact, the physiographic properties of the river systems act as filters to the highly variable precipitation input (Sutanto et al., 2020; Crochemore et al., 2020). Despite their low accuracy, WRI forecasts of irrigation needs seem to discriminate between below- and above-average conditions. Conversely, despite a seemingly high accuracy, forecasts of photovoltaic production reach low correlation levels, potentially due to irradiance being strongly dependent on cloud cover, which is unpredictable on time-scales longer than a few days. Extending the verification periods used for WRI and SEAP would however be necessary to confirm these performances. Lastly, both services applying the probabilistic verification framework reach reliability levels greater than 60% and can thus be considered acceptable to inform decision-making (cf. the cross-sector European-wide survey by Bruno-Soares et al. (2018) and the cross-sector worldwide serious game by Crochemore et al. (2021)).

The skill analysis (i.e. quality compared to that of benchmarks) provides a different perspective on the information provided by the services (Fig. 2). An increase in the quality of probabilistic forecasts with respect to the benchmarks requires an increase in sharpness without sacrificing reliability (Gneiting et al. 2007). Indeed, dynamical forecasts offer scenarios updated based on the latest ocean and atmosphere states and thus provide sharper ensembles than climatology-based forecasts, but sometimes at the cost of their reliability. Climatology-based ensembles offer perfect reliability, by nature, which services can, at best, equal, as in the case of SHYMAT (Fig. 2b). In the case of ROAT, where ensemble forecasts are sharper than climatology-based forecasts but less reliable, a trade-off between reliability and sharpness exists. This trade-off is well known and Weisheimer and Palmer (2014) have argued that ‘goodness’ should be first assessed in terms of reliability (or calibration) of the ensemble forecasts, since only reliable inputs may be considered for any forecast-based decision-making.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

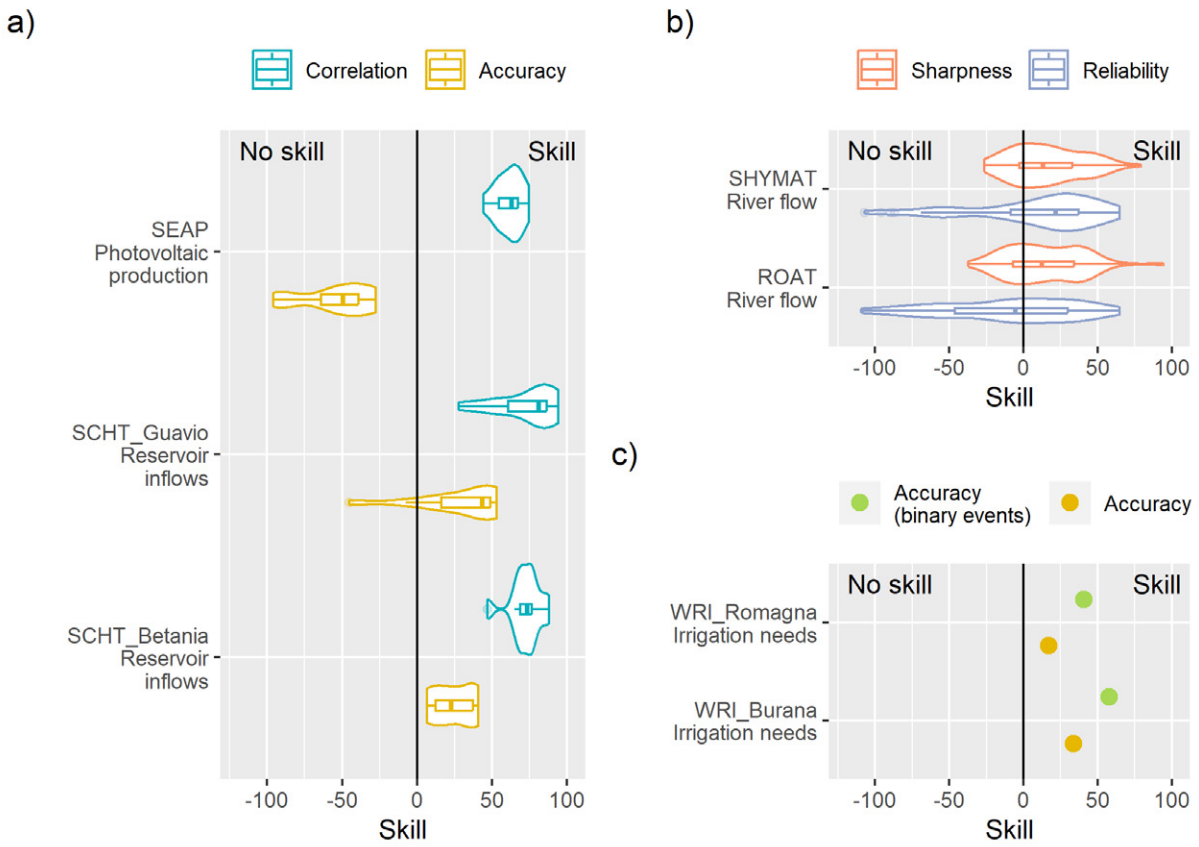

Fig. 2. Skill assessment of the services based on: a) deterministic, and b) probabilistic forecasts, and c) forecasts targetting threshold detection. The forecasts of the SCHT and WRI services are verified for two locations: the Guavio and Betania hydropower plants for SCHT, and Romagna and Burana for WRI. The black line indicates the skill value for which the service and benchmark forecasts have equivalent performances. Violins (showing the density) and boxplots (showing quartiles) indicate the variability in performance when the service supplies information at various time horizons (SHYMAT, ROAT) and/or initialisation months (SEAP, SCHT, SHYMAT, ROAT). See Table 2b '*Issue dates and horizons*' for more details.

The skill from deterministic forecasts varies with the service (Fig. 2a). SCHT improves upon the linear regression in terms of river flow correlation, which can be attributed to the AI-based forecasting method. Forecasts of photovoltaic production (SEAP) also show improved correlation compared to the benchmark despite low quality levels. In the case of SEAP, the high forecast accuracy does not translate in skill, showing that climatology-based approaches may be hard to beat due to the astronomical prevalence in determining solar production. In such cases, verifying the anomalies could reveal this earlier in the verification process.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

4. From forecast skill to service value

In ex-ante co-evaluation, the alternative knowledge source is based on expert knowledge rather than a prior service that does not exist yet. This explains why alternative knowledge sources (Table 2c) and benchmarks (Table 2b) do not coincide in the case of SCHT, SHYMAT and SEAP, and why skill (Fig. 2) and added value (Fig. 3) cannot be directly compared. This illustrates a challenge faced when trying to coordinate the verification and evaluation exercises at an early stage of service co-production, but that can be addressed in ex-post co-evaluation.

Nevertheless, in the case of ROAT, a link appears between the skill of inflows and the final service value in terms of reservoir volumes (Fig. 3). ROAT forecasts have skill in sharpness, but lose reliability with respect to climatology (Fig. 2b) leading to a final service value which is worse than business-as-usual in 33% of the cases (wet years). Yet, ROAT reaches a 39% median gain toward the value of a service with perfect information particularly in dry years (cf. Section 5). SHYMAT forecasts, however, gain in sharpness without sacrificing reliability (Fig. 2b) and their value is systematic and high (35% median gain). In the case of SCHT, the service value in both Guavio and Betania (27% and 6% median gains respectively) resembles their skill in accuracy (Fig. 2a). In Guavio, the forecast skill and service value are higher but more spread than in Betania, suggesting a possible link between skill and value despite discrepancies in benchmarks. A link between skill and value may thus be observed for services targeting the hydropower sector, potentially driven by more direct implications of forecast performance on economic gains in that sector (Cassagnole et al. 2021). Studies targetting reservoir operations and hydropower have further explored this relationship (Lee et al. 2021, Turner et al. 2017).

In supporting photovoltaic production, SEAP shows a systematic and high added-value with respect to the actual collectors' tracking method (78% median gain), when its photovoltaic production forecasts show skill in correlation and no skill in accuracy compared to photovoltaic production simulated with averaged historical data (Fig. 2a). This may suggest that forecast correlation prevails for such applications.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

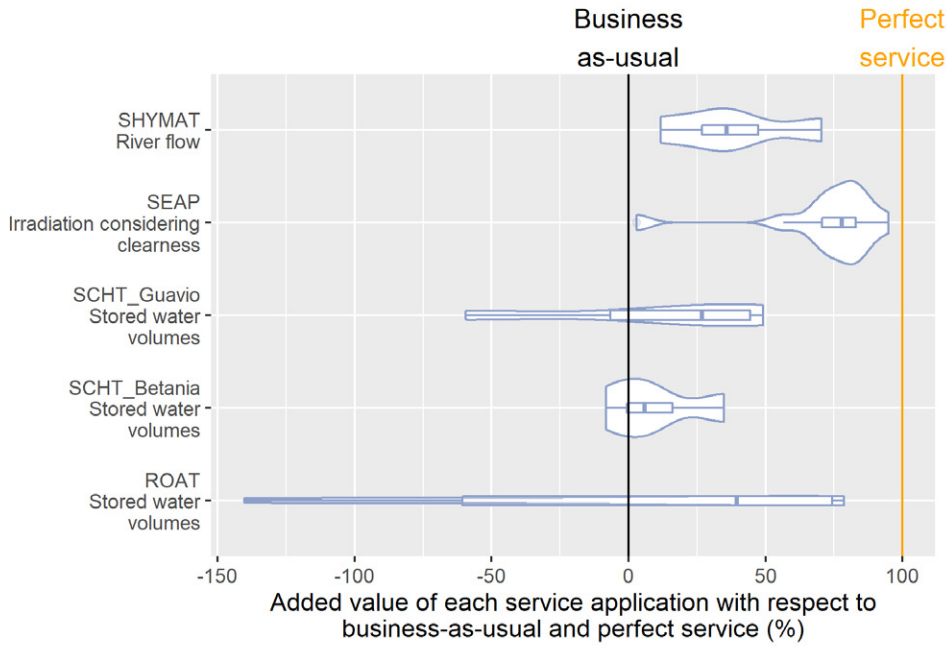

Fig. 3. Value of the services with respect to business-as-usual value (black line) and to a service with perfect information (orange line). Violins (showing the density) and boxplots (showing quartiles) indicate the variability in value when the service supplies information for various years (all but SEAP) and months. WRI does not appear because its value could not be assessed due to limited in-situ observations.

5. Lessons learnt through dialogues in the CLARA project

The dialogues between the service developers and users have highlighted several key messages. Firstly, iterations in co-evaluation played a role in shaping the services. For example, in SCHT the developers tested new algorithms to improve the inflow forecast skill, after the interactions between socio-economists and users, and a first round of evaluation. In the case of SHYMAT, these interactions led the developers to move from deterministic to probabilistic forecasts. A good representation of future uncertainties boosted service value in SHYMAT leading to higher informational benefits. Conversely, co-evaluation may be an opportunity to demonstrate the role of uncertainties, especially since their inclusion in the decision chains remains a challenge.

Service value varies depending on the hydro-climatic conditions. Developers of ROAT explained the wide interannual variability in value by considering dry and wet years separately. In wet years, the service underestimated water volumes causing spillway discharge that highly penalized payoffs and value, while drier conditions were better predicted leading to high payoffs nearly equal to that of a perfect service. Skill and value exercises should thus

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

demonstrate service potentials across hydro-climatic conditions to provide a comprehensive and nuanced picture. This retrospective exercise, however, requires long time series for robust analyses. Furthermore, service co-evaluation can inform investment strategies in multi-site optimization and management. The analysis of the SCHT value suggests that investments to improve the service in the Betania reservoir would be more beneficial than in Guavio, since the maximum potential value is higher in the former location.

Limitations in the availability of in-situ observations prevented a representative quantitative evaluation of the seasonal forecasts provided by WRI. This case thus illustrates a typical barrier to forecast verification and evaluation, which could be overcome over time by developing perennial measurement networks. Part of the data collecting process should thus ensure that the adequate observation networks are in place for such evaluations. In addition, short verification and evaluation periods lead to uncertain assessments, which should be interpreted with caution, could be quantified for instance through bootstrapping. Robust assessments thus require sustained efforts on data collection in time.

Lastly, quantitative evaluations only capture part of the final service value as the availability of the service alone may initiate changes in decision making. The Land Reclamation and Irrigation Consortium of Burana has recognized benefits from WRI, with farmers observing their water use as the service information about potential saves became available. On the contrary, the potential value of ROAT was high, but due to large penalties when relying on forecasts that were not fulfilled, the effective service value was very low. This is in line with the current manager's way of acting and perception of seasonal information. Managers apply an undeclared safety factor in their decisions and act more conservatively than stated in the value exercise. For this reason, it will be necessary to increase the reliability of the seasonal forecasts for the managers to include them in their daily workflow.

6. Recommendations for joint verification-evaluation

Our experience within CLARA was that implementing a joint verification-evaluation exercise while ensuring interdisciplinarity and transdisciplinarity was challenging. The project time allowed the co-development of the services, forecast verification, ex-ante co-evaluation, but not ex-post evaluation, and iterations were limited.

The verification and evaluation exercises took place over the three years of CLARA, but the interdisciplinary exchanges took place after. A comprehensive dialogue between disciplines (hydro-climate scientists and socio-economists) and actors (data purveyors, service

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

developers, service users) should occur early on so as to coordinate the verification and evaluation and ensure a seamless service assessment. This can be facilitated through participatory approaches. Time and anticipation are essential to homogenize terminologies, clarify disciplinary objectives and coordinate accordingly. In CLARA, multi-user forums involving developers and purveyors were organized every 6 months and allowed interactions, for instance through serious games targeting verification or business models. Such initiatives should outlive research projects, at least at the service scale, to allow for long-term and iterative co-evaluation.

Flexibility is required when screening methodologies to account for the needs and capacity of the users and developers. The co-evaluation process within CLARA was fully adapted to the service decision making contexts; verification, however, was partly designed to allow a cross-sector comparison between services. The comparison of both exercises revealed that the forecast variable being verified was not always the most relevant for users, the forecast benchmarks were sometimes unrealistic when forecasts were not already in place, and that the forecast attributes considered for the evaluation were not always assessed in verification. A perspective would be to design a verification better aligned with the evaluation methodology and hence with user needs, and propose metrics focusing on the forecast attributes leading to value.

Lastly, in climate services, business models highlight key scientific advancements that increase value against competitors (Larosa and Mysiak, 2020). The financial structure in particular (i.e. the costs and revenue architecture which was not considered here but analyzed within CLARA) can be used to secure the funding required for long-term and iterative co-evaluation.

7. The proposed co-development and co-evaluation framework

Based on this experience, we propose a framework that articulates well-known operational steps, namely *screening*, *collecting,* and *evaluating* or *verifying* (Fig. 4). The actors in the framework are categorized as social and climate scientists, service users and service developers, though these categories could overlap (e.g. climate scientists acting as service developers). The main recommendation being that actors in co-evaluation and co-development should interact early on in the exercise.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

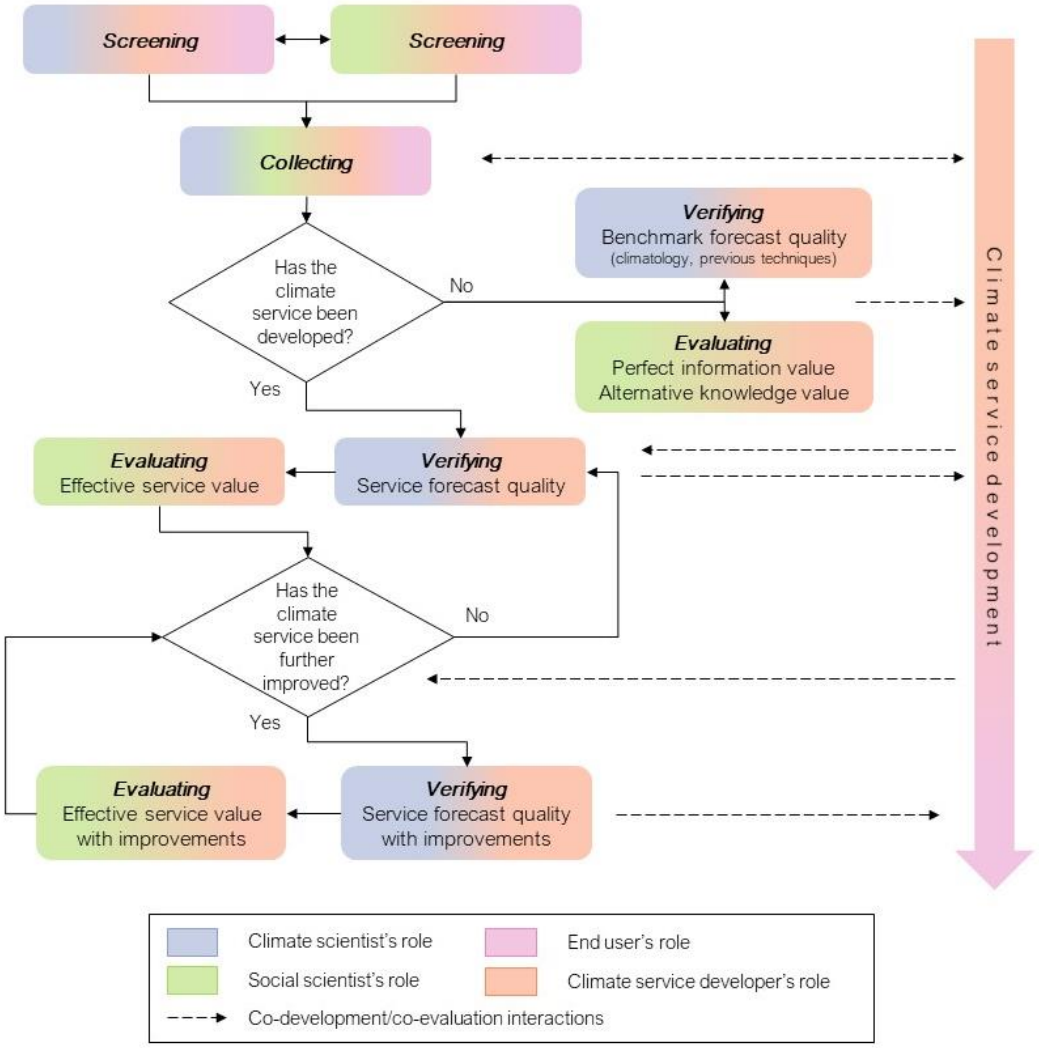

Fig. 4. Schematic representation of different phases of the forecast verification and service co-evaluation processes.

Screening consists in reviewing the methodologies to verify forecast quality or assess the economic value of climate services. Social and climate scientists actively discuss pros and cons of each method with users and developers in light of the climate services' features to determine which one can capture most of their salient characteristics.

Collecting involves all actors, namely climate and social scientists, climate services' producers and users to gather information to apply the proposed methodology. This is a truly transdisciplinary phase (Lawrence et al., 2022): the social and climate scientists indicate the type of information required, the users populate the exercise with the information related to the case study (when evaluating or verifying) and decision-making process (when evaluating), and the climate service producers provide the forecasts and other technical information related to

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

the service. This phase requires interactions between co-evaluation and co-development, since users discuss aspects which may change their requirements.

Evaluating and *verifying* are iterative steps: effective forecast quality and service value are assessed once the service is developed and every time there is an update or improvement. Improvements may originate from the quality of the seasonal forecasts, or the dissemination and communication means. The updated service verification and evaluation thus inform on the improvements worth pursuing (e.g. advanced modeling techniques, data assimilation, uncertainty communication).

Verification aims to assess forecast quality, which is process-, system-, location-, and time-dependent. The verification process is generally independent from the decision-making process, though it should ideally be designed to assess forecast attributes of relevance for the users. The quality of a forecast set is commonly assessed against a *reference,* e.g. observations from a long enough past period (or simulations in the absence of observations) to ensure robust quality estimates. It is then compared with a forecast *benchmark*, which either represents the best available forecast prior to the development of the service or a prior version of the service (Pappenberger et al., 2015; Wilks, 2011).

When *evaluating* a climate service, value can be generated in different steps of the value chain, and thus may require different evaluation techniques (Perrels et al., 2013, Fig. 5). For service prototypes aiming to support operational decision planning; value originates not only from the quality of the provided forecasts, but also from how well the information is tailored (for instance the choice of the metrics presenting the economic service value) and communicated (Calvo et al., 2022), and how well the service supports decision-making and user needs (Materia et al., 2020). A quantitative evaluation as performed in CLARA could thus be complemented by a qualitative evaluation accounting for value drivers other than the forecasts.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

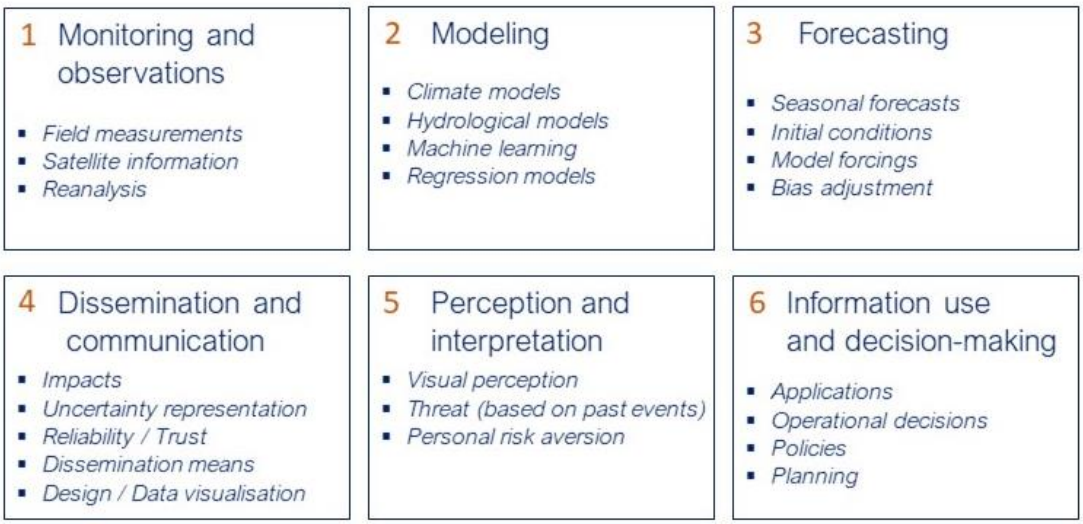

Fig. 5. A representative value chain for climate services, from monitoring and observations down to decision-making. Each step is identified and chained with upstream and downstream steps, while examples of the methods or components involved in the investigated services are provided below each step.

8. Conclusions

The joint verification and evaluation exercise performed within CLARA across services highlights the diversity of situations that can be encountered when deploying the joint framework to support co-production. The plurality in service designs reflects the plurality of decision-making contexts and user needs, which will also influence the skill-value relationship. Previous works (e.g. Cloke et al., 2017; Materia et al., 2020) suggest that improvements in early warning systems shall be assessed considering both skill and value. This work further suggests that (a) choices in the co-development of climate services should consider both forecast skill and service value, (b) time and, therefore, sustained funding are key to allow for such interdisciplinary and transdisciplinary exercises, and (c) verification exercises should evolve to target user needs in diverse sector-related decision-making contexts.

Acknowledgments.

This research has been conducted with support from the European Union's Horizon 2020 project CLARA (Climate forecast enabled knowledge services) under Grant Agreement 730482. L. Crochemore activities fit within the PEPR TRACCS (Transform climate modelling for climate services) funded by the ANR (French National Research Agency). S. Materia has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101033654. I.G. Pechlivanidis was funded by the EU Horizon 2020 projects I-CISK (Innovating climate services through

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

integrating scientific and local knowledge) under Grant Agreement 101037293, and CLINT (Climate Intelligence: Extreme events detection, attribution and adaptation design using machine learning) under Grant Agreement 101003876.

Data Availability Statement.

The seasonal forecasts used in the services are all freely available via Copernicus Climate Change Service. The services are meant for commercialisation, and output data are thus sensitive, but some of the service platforms are openly accessible, such as SEAP, SHYMAT (paid access for access to some run-of-river systems), ROAT (paid access for access to some reservoir systems), WRI (login required), SCHT's demo versions (restricted to non-sensitive data).

APPENDIX A

Formulation of forecast attributes and skill

In the following equations, the notations are as follows : N is the number of years y, m is the forecast month and lt the forecast lead time. All metrics are then normalized so that their optimum value is 100.

A.1 Accuracy

The accuracy is measured with the mean absolute error, formulated as:

$$MAE_{m,lt} = \frac{1}{N}\sum_{y=1}^{N}\left|F_{y,m,lt} - O_{y,m,lt}\right| \qquad \text{(A.1)}$$

where $F_{y,m,lt}$ is the deterministic forecast or the mean (SCHT, SEAP) or median (WRI) of the ensemble forecast, and $O_{y,m,lt}$ is the reference observations for year y and month $m+lt-1$.

A.2 Correlation

The correlation of forecasts with the reference is assessed based on the Spearman rank correlation coefficient. It is then multiplied by a factor of 100.

A.3 Reliability

Reliability is a characteristic of probabilistic forecasts that indicates the consistency between observed and forecast probabilities. In the literature, reliability is commonly based on a division of an ensemble forecast range into bins delimited by the ensemble members. Reliability can then be assessed based on how evenly the retrospective observation falls within

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

each of these bins, for instance through a rank histogram (Anderson, 1996). Here, instead, the forecast is simplified to a single bin comprising the entire range between the 10th and 90th quantiles of the forecast ensemble for each time step (a single combination of year *y*, month *m* and lead time *lt*). The reliability *R*, in this paper, is then formulated as the frequency with which the observation falls within this bin:

$$R_{m,lt} = \frac{100}{N}\sum_{y=1}^{N}\delta_{y,m,lt} \qquad \text{(A.2)}$$

where

$$\begin{cases}\delta_{y,m,lt} = 1 \; if \; O_{y,m,lt} \in \left[Q_{10}\left(F_{y,m,lt}\right), Q_{90}\left(F_{y,m,lt}\right)\right] \\ \delta_{y,m,lt} = 0 \; otherwise\end{cases}$$

A.4 Sharpness

Sharpness indicates the spread of the ensemble forecast members with respect to the spread in historical observations (Wilks, 2011). Here, it is expressed as the ratio between the 80% quantile range of the service-generated forecast (difference between the 90th and 10th percentiles of the forecast ensemble) against the 80% quantile range of climatology:

$$S_{m,lt} = \frac{\sum_{y=1}^{N}\left(Q_{90}(F_{y,m,lt}) - Q_{10}(F_{y,m,lt})\right)}{\sum_{y=1}^{N}\left(Q_{90}(C_{y,m,lt}) - Q_{10}(C_{y,m,lt})\right)} \qquad \text{(A.3)}$$

where $F_{y,m,lt}$ is the ensemble forecast, and $C_{y,m,lt}$ is the ensemble based on historical observations for month *m* and lead time *lt*, excluding year *y*. It is independent from observations and, therefore, is not a quality indicator per se. Here, we consider it as complementary to the reliability indicator since "Sharp forecasts will be accurate only if they also exhibit good reliability" (Wilks, 2011).

A.5 Accuracy in the case of binary events

The capacity of the ensemble forecasts to predict whether an event will occur or not (binary event) is evaluated based on the Brier score (Brier, 1950). The occurrence of an event is defined based on the probability of reaching a fixed threshold. It is formulated as follows:

$$Brier_{m,lt} = \frac{1}{N}\sum_{y=1}^{N}\left(P_{F,y,m,lt} - P_{O,y,m,lt}\right) \qquad \text{(A.4)}$$

where $P_{F,y,m,lt}$ is the forecast probability of the event occurring (between 0 and 1) and $P_{O,y,m,lt}$ is 1 if the event occurred and 0 otherwise.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

A.6 Skill score

Skill scores compare any forecast characteristics or quality indicator to that of a reference forecast. Its most common formulation (Wilks, 2011; Jolliffe and Stephenson, 2003), ranging from 100 (optimum) to infinite negative values, is as follows:

$$Skill_{m,lt} = \frac{Crit_{m,lt}^{CS} - Crit_{m,lt}^{BaU}}{Crit_{m,lt}^{PI} - Crit_{m,lt}^{BaU}} \times 100\% \qquad \text{(A.5)}$$

where $Crit_{m,lt}^{CS}$ is the quality of the service-generated forecast, $Crit_{m,lt}^{PI}$ is the quality of a perfect forecast system and $Crit_{m,lt}^{BaU}$ the quality of the reference forecast. The skill score in forecast verification relates closely to the performance index in service evaluation (see Appendix B).

APPENDIX B

Operationalization of the value assessment methodology

Applying the value of information theory (Pope et al. 2017, Pope et al. 2019) requires inputs on the decision-making context through an *indicator*, a *decision space*, *states of the world*, *payoffs*, and forecast inputs of varying *skill*.

The *indicator* represents the link between the forecasts and the operative decisions. It could be the variable forecast by the service (as river flow in m3/s for SHYMAT) or a variable derived from the forecast. For SCHT, the service forecasts inflows to the reservoir, but the decision is taken based on the water volume in the reservoir. A reservoir management model was thus made available by the user to derive reservoir volumes from past inflows given reservoir characteristics. Similarly, SEAP produces forecasts of photovoltaic production, but states of the world are based on the daily irradiation on photovoltaic modules based on the clearness index kt.

The *decision space* translates the user operative decisions. Since these services are mock-ups and decisions are not already in place, this decision space represents the users' decision to be supported once the service becomes operational. The decision space could consist of a "do-don't" decision, or be continuous as in the case of SHYMAT and SCHT (setting the turbine level). Similarly, SEAP allows solar trackers to be programmed in advance, thus the action is to position the solar trackers depending on the class of the clearness index. For ROAT the decision on water management leads to three actions: "save water", "maintain water level in the dam", and "release water".

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

The user decision is expected to change depending on the *state of the world* the indicator falls into. These states can be constant in time (SCHT, SHYMAT) or change monthly (ROAT). For SCHT, the states of the world derive from the management rules of the dam and affect the level of turbine activation; for SHYMAT, instead, the states of the world derive from different levels of river flows, that ultimately affect the operation of the turbines. In the case of ROAT, some states are constant in time (as the flood control states); others change monthly based on the dam operating rules and the river basin management plan.

The *payoffs* associated with each action-state of the world combination can be presented in terms of a monetary or a physical unit of measure, or a different unit. In some cases, the value of the physical unit is unknown and the best way to proceed is to use a physical unit of measure acting as a proxy for savings or profits (as in SEAP). Because the presented services are prototypes and users were not always allowed to disclose potential profits, some users were asked to provide instead a score for combinations of action-forecasts-effective realizations. The final value was then given by the score. This methodology is applied for ROAT, SHYMAT and SCHT. ROAT's users define payoffs as linearly increasing inside the "normal state" of the world (the higher the water level the higher the payoff), and as discrete in the other states of the world. Moreover, since the service works in a multipurpose context (multi-objective reservoir), payoffs are set by three different users before being weighed. SHYMAT and SCHT users consider payoffs related to operational management only. However, they recognize that climate services could be used to assess, for instance, potential financial costs for the producer when entering the energy market or covering missing revenues by accessing credit markets (SCHT) or to start machine repairs (SHYMAT).

Finally, the *performance* of the knowledge stocks represents the ability of a knowledge source to forecast states of the world. It relates to the quality assessment presented in this paper but considers which state of the world the hindcast predicts. The performance here consists of two contingency matrices, one for each knowledge stock. In this case, we are interested in the "right" forecasts (true negatives and true positives), the "wrong" ones (false positives and false negatives) and in overestimations and underestimations, since each combination is linked to a potential course of action and associated payoffs.

The expected value is assessed for each knowledge stock, that is for the business-as-usual knowledge (BaU) and the climate service knowledge (CS):

$$EVBaU = \sum_{k=1}^{N} p_{BaU}(X_k)\, min_A \sum_{n=1}^{N} \pi(A, x_n) P_{BaU}\,(x_n|X_k) \qquad \text{(B.1)}$$

$$EVCS = \sum_{k=1}^{N} p_{CS}(X_k)\, min_A \sum_{n=1}^{N} \pi(A, x_n) P_{CS}\,(x_n|X_k) \qquad \text{(B.2)}$$

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

Where $p(X_k)$ is the forecast probability of event X_k according to the knowledge source (BaU or CS), $\pi(A, x_n)$ is the payoff associated with action A when the state of the world x_n occurs, $P(x_n|X_k)$ is the probability of event x_n being observed when X_k is forecast.

The Expected Value of Perfect Information (*EVPI*) (Pope et al. 2017) is derived from the effective realizations in the test period:

$$EVPI = \sum_{n=1}^{N} min_A \pi(A, x_n) p(X_n) - EVBaU \qquad \text{(B.3)}$$

also called "maximum potential value". This is the upper bound for the climate service value. Finally, we calculate the effective value of the climate service (or Expected Value of Imperfect Information; *EVII*) as:

$$EVII = EVCS - EVBaU \qquad \text{(B.4)}$$

Moreover, following Richardson (2000), in each iterative phase of the evaluation when new improvements in the climate service are assessed we calculate a performance index (*PI*) as:

$$PI = EVII/EVPI \qquad \text{(B.5)}$$

PI captures how much of the maximum potential value is effectively gained due to the development level of the climate service. *PI* is equivalent to the concept of forecast skill used in forecast verification (see Appendix A).

SIDEBAR

Verification and evaluation terminologies

Terminologies are sometimes field-dependent. We thus define the terminologies used throughout the manuscript:

- *Forecast quality:* A measure of how correctly the forecast corresponds to the observations or simulations in the absence of observations (Murphy, 1993)
- *Skill score:* A measure of the relative quality of a forecast set, assessed with respect to some reference forecasts, such as forecast climatology, persistence, random forecast (Wilks, 2011)
- *Service value* : Impact on user benefits of using climate information compared to using another knowledge source in decision-making (Meza et al. 2008)

Appendix A details the formulations of forecast attributes and skill, and Appendix B the formulations of service value.

REFERENCES

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

Alexander, M., and S. Dessai, 2019: What can climate services learn from the broader services literature? *Climatic Change*, **157**, 133–149, https://doi.org/10.1007/s10584-019-02388-8.

Anderson, J. L., 1996: A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations. Journal of climate, 9, 1518–1530, https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

Batté, L., L. Dorel, C. Ardilouze, and J.-F. Guérémy, 2019: *Documentation of the METEO-FRANCE seasonal forecasting system 7*. Météo-France / ECMWF, http://www.umr-cnrm.fr/IMG/pdf/system7-technical.pdf (Accessed October 13, 2021).

Boon, E., S. J. Wright, R. Biesbroek, H. Goosen, and F. Ludwig, 2022: Successful climate services for adaptation: What we know, don't know and need to know. *Climate Services*, **27**, 100314, https://doi.org/10.1016/j.cliser.2022.100314.

Brasseur, G. P., and L. Gallardo, 2016: Climate services: Lessons learned and future prospects. *Earth's Future*, **4**, 79–89, https://doi.org/10.1002/2015EF000338.

Bremer, S., A. Wardekker, S. Dessai, S. Sobolowski, R. Slaattelid, and J. van der Sluijs, 2019: Toward a multi-faceted conception of co-production of climate services. *Climate Services*, **13**, 42–50, https://doi.org/10.1016/j.cliser.2019.01.003.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Bruno Soares, M., M. Alexander, and S. Dessai, 2018: Sectoral use of climate information in Europe: A synoptic overview. *Climate services*, **9**, 5–20, https://doi.org/10.1016/j.cliser.2017.06.001.

Calvo, L., I. Christel, M. Terrado, F. Cucchietti, and M. Pérez-Montoro, 2022: Users' Cognitive Load: A Key Aspect to Successfully Communicate Visual Climate Information. Bulletin of the American Meteorological Society, 103, E1–E16, https://doi.org/10.1175/BAMS-D-20-0166.1.

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, (Accepted)*.

Cassagnole, M., M.-H. Ramos, I. Zalachori, G. Thirel, R. Garçon, J. Gailhard, and T. Ouillon, 2021: Impact of the quality of hydrological forecasts on the management and revenue of hydroelectric reservoirs – a conceptual approach. *Hydrology and Earth System Sciences*, **25**, 1033–1052, https://doi.org/10.5194/hess-25-1033-2021.

Clements, J., A. Ray, and G. Anderson, 2013: *The Value of Climate Services across Economic and Public Sectors: A Review of Relevant Literature*. United States Agency for International

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0026.1.

Development, https://www.climatelinks.org/sites/default/files/asset/document/CCRD-Climate-Services-Value-Report_FINAL_0.pdf (Accessed September 1, 2021).

Cloke, H. L., F. Pappenberger, P. J. Smith, and F. Wetterhall, 2017: How do I know if I've improved my continental scale flood early warning system? *Environmental Research Letters*, **12**, 044006, https://doi.org/10.1088/1748-9326/aa625a.

Contreras, E., J. Herrero, L. Crochemore, C. Aguilar, and M. J. Polo, 2020a: Seasonal Climate Forecast Skill Assessment for the Management of Water Resources in a Run of River Hydropower System in the Poqueira River (Southern Spain). *Water*, **12**, 2119, https://doi.org/10.3390/w12082119.

Contreras, E., J. Herrero, L. Crochemore, I. Pechlivanidis, C. Photiadou, C. Aguilar, and M. J. Polo, 2020b: Advances in the Definition of Needs and Specifications for a Climate Service Tool Aimed at Small Hydropower Plants' Operation and Management. *Energies*, **13**, 1827, https://doi.org/10.3390/en13071827.

Crochemore, L., C. Cantone, I. G. Pechlivanidis, and C. S. Photiadou, 2021: How does seasonal forecast performance influence decision-making? Insights from a serious game. *Bulletin of the American Meteorological Society*, 1–54, https://doi.org/10.1175/BAMS-D-20-0169.1.

Crochemore, L., M.-H. Ramos, and I. G. Pechlivanidis, 2020: Can Continental Models Convey Useful Seasonal Hydrologic Information at the Catchment Scale? *Water Resources Research*, **56**, e2019WR025700, https://doi.org/10.1029/2019WR025700.

Delpiazzo, E., F. Bosello, P. Mazzoli, S. Bagli, V. Luzzi, and F. Dalla Valle, 2022: Co-evaluation of climate services. A case study for hydropower generation. *Climate Services*, **28**, 100335, https://doi.org/10.1016/j.cliser.2022.100335.

Essenfelder, A. H., and Coauthors, 2020: Smart Climate Hydropower Tool: A Machine-Learning Seasonal Forecasting Climate Service to Support Cost–Benefit Analysis of Reservoir Management. *Atmosphere*, **11**, https://doi.org/10.3390/atmos11121305.

Giuliani, M., Crochemore, L., Pechlivanidis, I., and Castelletti, A., 2020: From skill to value: isolating the influence of end user behavior on seasonal forecast assessment, Hydrol. Earth Syst. Sci., **24**, 5891–5902, https://doi.org/10.5194/hess-24-5891-2020.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69, 243–268.

Hargreaves, J. C., 2010: Skill and uncertainty in climate models. *WIREs Climate Change*, **1**, 556–564, https://doi.org/10.1002/wcc.58.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **n/a**, https://doi.org/10.1002/qj.3803.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0026.1.

Johnson, S. J., and Coauthors, 2019: SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, **12**, 1087–1117, https://doi.org/10.5194/gmd-12-1087-2019.

Jolliffe, I. T., and D. B. Stephenson, 2003: Forecast Verification: A Practitioner's Guide in Atmospheric Science. John Wiley & Sons Ltd., 240 pp.

Katz, R. W., and J. Lazo, 2011: Economic Value of Weather and Climate Forecasts. *The Oxford Handbook of Economic Forecasting*, https://doi.org/10.1093/oxfordhb/9780195398649.013.0021.

Katz, R. W., and A. H. Murphy, eds., 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, https://doi.org/10.1017/CBO9780511608278.

Larosa, F., and J. Mysiak, 2020: Business models for climate services: An analysis. *Climate Services*, **17**, 100111, https://doi.org/10.1016/j.cliser.2019.100111.

Laugesen, R., M. Thyer, D. McInerney, and D. Kavetski, 2023: Flexible forecast value metric suitable for a wide range of decisions: application using probabilistic subseasonal streamflow forecasts. *Hydrology and Earth System Sciences*, 27, 873–893, https://doi.org/10.5194/hess-27-873-2023.

Lawrence, M. G., S. Williams, P. Nanz, and O. Renn, 2022: Characteristics, potentials, and challenges of transdisciplinary research. *One Earth*, **5**, 44–61, https://doi.org/10.1016/j.oneear.2021.12.010.

Lee, D., J. Y. Ng, S. Galelli, and P. Block, 2021: Unfolding the relationship between seasonal forecast skill and value in hydropower production: A global analysis. *Hydrology and Earth System Sciences Discussions*, **2021**, 1–27, https://doi.org/10.5194/hess-2021-518.

Materia, S., Á. G. Muñoz, M. C. Álvarez-Castro, S. J. Mason, F. Vitart, and S. Gualdi, 2020: Multimodel Subseasonal Forecasts of Spring Cold Spells: Potential Value for the Hazelnut Agribusiness. *Weather and Forecasting*, **35**, 237–254, https://doi.org/10.1175/WAF-D-19-0086.1.

Mauser, W., G. Klepper, M. Rice, B. S. Schmalzbauer, H. Hackmann, R. Leemans, and H. Moore, 2013: Transdisciplinary global change research: the co-creation of knowledge for sustainability. *Current Opinion in Environmental Sustainability*, **5**, 420–431, https://doi.org/10.1016/j.cosust.2013.07.001.

Meza, F. J., J. W. Hansen, and D. Osgood, 2008: Economic Value of Seasonal Climate Forecasts for Agriculture. Journal of Applied Meteorology and Climatology, 47, 1269–1286.

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.

Pappenberger, F., M. H. Ramos, H. L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, and P. Salamon, 2015: How do I know if my forecasts are better? Using benchmarks in

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.

hydrological ensemble prediction. *Journal of Hydrology*, **522**, 697–713, http://dx.doi.org/10.1016/j.jhydrol.2015.01.024.

Pechlivanidis, I. G., Crochemore, L., Rosberg, J., & Bosshard, T., 2020: What are the key drivers controlling the quality of seasonal streamflow forecasts? Water Resources Research, **56**, e2019WR026987. https://doi.org/10.1029/2019wr026987

Perrels, A., Th. Frei, F. Espejo, L. Jamin, and A. Thomalla, 2013: Socio-economic benefits of weather and climate services in Europe. *Advances in Science and Research*, **10**, 65–70, https://doi.org/10.5194/asr-10-65-2013.

Photiadou, C., and Coauthors, 2021: Designing a Climate Service for Planning Climate Actions in Vulnerable Countries. *Atmosphere*, **12**, https://doi.org/10.3390/atmos12010121.

Pope, E. C. D., C. Buontempo, and T. Economou, 2017: Quantifying how user-interaction can modify the perception of the value of climate information: A Bayesian approach. Climate Services, 6, 41–47, https://doi.org/10.1016/j.cliser.2017.06.006.

——, ——, and ——, 2019: Exploring constraints on the realised value of a forecast-based climate service. Climate Services, 15, 100102, https://doi.org/10.1016/j.cliser.2019.100102.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. Quarterly Journal of the Royal Meteorological Society, 126, 649–667, https://doi.org/10.1002/qj.49712656313.

Sanna, A., A. Borrelli, P. Athanasiadis, S. Materia, A. Storto, A. Navarra, S. Tibaldi, and S. Gualdi, 2016: *CMCC-SPS3: The CMCC Seasonal Prediction System 3*. CMCC, https://www.cmcc.it/publications/rp0285-cmcc-sps3-the-cmcc-seasonal-prediction-system-3 (Accessed October 13, 2021).

Street, R. B., C. Buontempo, J. Mysiak, E. Karali, M. Pulquério, V. Murray, and R. Swart, 2019: How could climate services support disaster risk reduction in the 21st century. *International Journal of Disaster Risk Reduction*, **34**, 28–33, https://doi.org/10.1016/j.ijdrr.2018.12.001.

Sutanto, S. J., F. Wetterhall, and H. A. J. Van Lanen, 2020: Hydrological drought forecasts outperform meteorological drought forecasts. Environmental Research Letters, 15, 084010, https://doi.org/10.1088/1748-9326/ab8b13.

Troccoli, A., 2018: Achieving Valuable Weather and Climate Services. *Weather & Climate Services for the Energy Industry*, A. Troccoli, Ed., Springer International Publishing, 13–25.

Turner, S. W. D., J. C. Bennett, D. E. Robertson, and S. Galelli, 2017: Complex relationship between seasonal streamflow forecast skill and value in reservoir operations. *Hydrology and Earth System Sciences*, **21**, 4841–4859, https://doi.org/10.5194/hess-21-4841-2017.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0026.1.

Vaughan, and Dessai, 2014: Climate services for society: origins, institutional arrangements, and design elements for an evaluation framework. *Wiley Interdisciplinary Reviews: Climate Change*, **5**, 587–603, https://doi.org/10.1002/wcc.290.

Vaughan, C., M. F. Muth, and D. P. Brown, 2019: Evaluation of regional climate services: Learning from seasonal-scale examples across the Americas. *Climate Services*, **15**, 100104, https://doi.org/10.1016/j.cliser.2019.100104.

Villani, G., F. Tomei, V. Pavan, A. Pirola, A. Spisni, and V. Marletto, 2021: The iCOLT climate service: Seasonal predictions of irrigation for Emilia-Romagna, Italy. *Meteorological Applications*, **28**, e2007, https://doi.org/10.1002/met.2007.

Weisheimer, A., and T. N. Palmer, 2014: On the reliability of seasonal climate forecasts. *Journal of The Royal Society Interface*, **11**, 20131162, https://doi.org/10.1098/rsif.2013.1162.

Wilks, D. S. (2011). Statistical methods in the atmospheric sciences (Vol. 100). Academic press. Ed., Vol. 100 of, Academic Press, 301–394.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0026.1.